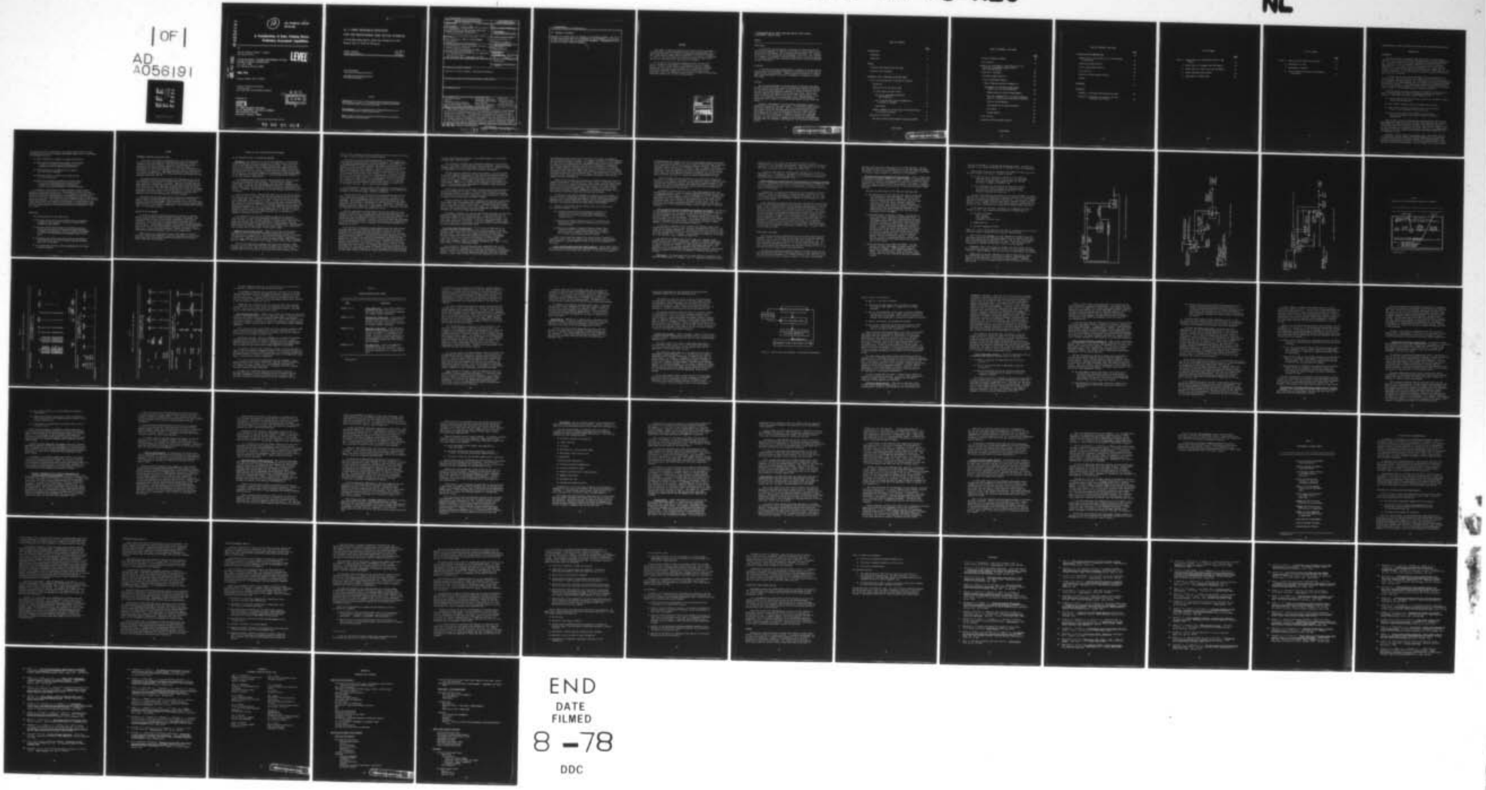
AD-A056 191 LITTON SYSTEMS INC FORT BENNING GA MELLONICS SYSTEMS--ETC F/G 5/9
A CONSIDERATION OF ARMY TRAINING DEVICE PROFICIENCY ASSESSMENT --ETC(U)
JUN 78 J B SHELNUTT, R J SMILLIE, J BERCOS DAHC19-77-C-0011

UNCLASSIFIED ARI-TR-78-A20 NL



| OF |
AD
A056191

END
DATE
FILMED
8 -78
DDC

12

2

ARI TECHNICAL REPORT
TR-78-A20

A Consideration of Army Training Device Proficiency Assessment Capabilities

by

Jack B. Shelnutt, Robert J. Smillie,
and James Bercos

LITTON MELLONICS SYSTEMS DEVELOPMENT DIVISION
DEFENSE SCIENCES LABORATORIES
P. O. Box 2498
Fort Benning, Georgia 31905

new? name change?

JUNE 1978

410654

Contract DAHC 19-77-C-0011

Dorothy L. Finley, Work Unit Leader
H. C. Strasel, Chief
ARI FIELD UNIT, FORT BENNING, GEORGIA

Prepared for

ari

U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
5001 Eisenhower Avenue
Alexandria, Virginia 22333

aett

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the
Deputy Chief of Staff for Personnel

JOSEPH ZEIDNER
Acting Technical Director

W. C. MAUS
COL, GS
Commander

Research accomplished
under contract to the Department of the Army

Litton Mellonics Systems Development Division
Defense Sciences Laboratories

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

19

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
TR-78-A20 18 ARI

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER
9

4. TITLE (and Subtitle) 6
A CONSIDERATION OF ARMY TRAINING DEVICE PROFICIENCY ASSESSMENT CAPABILITIES.

5. TYPE OF REPORT & PERIOD COVERED
Technical rept.,
~~TASK REPORT~~

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s) 10
Jack B./Shelnutt, Robert J./Smillie ~~and~~
James/Bercos

8. CONTRACT OR GRANT NUMBER(s) 15
DAHC 19-77-C-0011

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Litton-Mellonics System Development Division
Defense Sciences Laboratory
P.O. Box 2498, Fort Benning, GA 31905

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 16
2Q762722A765

11. CONTROLLING OFFICE NAME AND ADDRESS
U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Ave., Alexandria, VA 22333

12. REPORT DATE 11
June 1978

13. NUMBER OF PAGES
65 12 73 p.

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Cost Effectiveness
Fidelity
Information Needs Analysis
Initial Measurement Analysis
Operational Readiness Assessment
Performance Evaluation
Performance Measurement
Pilot Testing
Proficiency Assessment
Proficiency Testing
Simulators
System Cost Effectiveness Testing
Training Devices
Training Effectiveness

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report reviews the procedures and problems involved in the assessment of the use of training devices as a cost-effective alternative to the use of operational equipment for the evaluation of individual and collective proficiency in the U.S. Army. A review of the literature was conducted as well as an informal survey of personnel in other agencies who are involved in the use of training devices for proficiency assessment. This information was employed to: (a) review the use of training devices in proficiency assessment programs by agencies other than the Army; (b) to summarize aspects of proficiency test

CONT'

20. ABSTRACT (continued)

programs in the Army which are relevant to the present problem; and (c) to discuss issues which need to be considered in the assessment of the utility of using training devices for proficiency assessment. Recommendations were provided for future research planning.

FOREWORD

This report is one of a series on the research support provided by the Mellonics Systems Development Division of Litton Systems, Inc., to the Army Research Institute for the Behavioral and Social Sciences (ARI) under Contract Number DAHC 19-77-C-0011. The report, as submitted, is a part of the final report of the total contractual support effort and will be incorporated into that report by reference.

As set forth in the Contract Statement of Work, the Mellonics effort includes the performance of investigatory studies in human factors. As part of that effort, this report reviews the procedures and problems that need to be considered in an investigation of the proficiency assessment capabilities of Army training devices and establishes a basis for designing future studies to assess the proficiency assessment capabilities of Army training devices.

A CONSIDERATION OF ARMY TRAINING DEVICE PROFICIENCY ASSESSMENT CAPABILITIES

BRIEF

Requirement:

To review the state-of-the-art in the use of training devices for proficiency assessment, to identify the evaluation system context within the Army unit environment, and to evaluate the requirements for research on the use of training devices for assessment in the unit. This research is in support of a program to further develop the use of training devices for training assessment and, especially, operational readiness assessment in Army field units.

Procedure:

A literature review was conducted and a telephone survey was made of military and civilian personnel cognizant in the areas of training device usage and test and evaluation. A model of decision processes in measurement system development was used as a framework to integrate the identified issues.

Findings:

The use and development of training device performance measurement systems, especially for the purpose of operational readiness evaluations, is a relatively undeveloped area. The use of device measurement systems for both training and readiness assessment is primarily found in the aviation community and appears to be an emerging, albeit inconsistent, trend in both military and civilian areas. The inconsistencies in use result from differences in organizational policy, command emphasis, device measurement capability, and device measurement system upkeep and support.

Systematic development of training device measurement systems is suggested to consist of four major steps: (1) information needs analysis, (2) initial measurement analysis, (3) pilot testing and selection of final measure sets, and (4) measurement system effectiveness testing. The consensus of the literature and expert opinion is that training device measurement systems are infrequently systematically developed, that this is a limiting factor in the use of devices for assessment purposes, and that research is needed to better develop the methodology and data bases for systematic development of training device measurement systems.

TABLE OF CONTENTS

(continued)

TABLE OF CONTENTS (continued)

(continued)

TABLE OF CONTENTS (concluded)

LIST OF TABLES

LIST OF FIGURES

A CONSIDERATION OF ARMY TRAINING DEVICE PROFICIENCY ASSESSMENT CAPABILITIES

INTRODUCTION

BACKGROUND

Current estimates of allowable post mobilization training periods indicate a "come as you are" war for the majority of Army units (Reference 22). The rapid response requirement necessitates that operational units must continually train to maintain a satisfactory level of readiness (Reference 67). The need to maintain a high proficiency level through training in operational units will be accompanied by a related need for frequent evaluations of individual and unit proficiency (References 23, 67) to verify proficiency levels and redirect training if requisite proficiency levels are not maintained.

In recognition of the need for improved training and evaluation at the unit level, the Army has begun to systematize its training and evaluation programs (References 39, 61, 67). The adoption of the systems approach has produced an increased emphasis on job-relevant and performance-oriented training and evaluation. This emphasis has created even greater demands for the use of operational equipment in training and in evaluation and a concomitant increase in the projected costs of training evaluation.

Training devices are being increasingly employed by the Army in unit-level training programs to reduce the demand for operational systems because in comparison with those systems (Reference 35):

- training devices usually have lower acquisition, operation, support and/or maintenance costs;
- they normally provide a safer environment for training;
- they usually create less demand on limited unit resources such as time, personnel, or facilities; and
- they are often simply a more effective and efficient learning environment due to their unique instructional capabilities.

Some of these current and near-future training devices have performance measurement capabilities that not only allow monitoring of training performance, but also provide the additional cost-effective benefit of furnishing a vehicle for proficiency testing. Most devices, however, have not been designed with proficiency testing in mind and as a consequence do not have the necessary objective performance measurement capability built in (References 4, 62).

Research in several areas of psychology (*e.g.*, proficiency testing, training device design, complex performance measurement) indicates that it would be beneficial for training devices to be developed with instrumentation and programmed procedures designed to provide reliable, valid, and useful measures of performance. This literature also raises a number of issues that must be considered in the design and assessment of the effectiveness of these performance measurement capabilities. Chief among the considerations are:

- What information is needed to evaluate proficiency?
- What design condiderations must apply to the devices in order to provide useful information that is accepted as valid by operational personnel?
- What determines the acceptability of a device for proficiency testing?
- Which cost model is correct for costing the use of training devices?
- What are the appropriate measures of effectiveness to be used in determining the relative utility of proficiency data generated in operational systems versus data generated in training devices?

A systematic investigation of these and other related issues is currently being planned by the U. S. Army Research Institute for the Behavioral and Social Sciences (ARI). This research will be specifically devoted to developing methodologies and procedures for evaluatinq the utility of using unit training devices as an assessment tool in Army proficiency test programs. This report provides a partial foundation for that ARI effort and discusses the issues to be considered in designing and using training devices as cost-effective alternatives to the operational equipment for the assessment of individual and collective proficiency in the U. S. Army.

OBJECTIVES

The four objectives of this report are:

- To review current uses of training devices for evaluation purposes by other agencies in order to identify their procedures and problems.
- To review current Army programs of individual and crew proficiency evaluation in the operation and maintenance of weapons and equipment in order to identify the context for use of training devices as assessment tools in the Army.
- To identify and review issues which must be considered in the evaluation of the use of training devices as assessment tools in the Army.
- To use the above reviews to make recommendations for future research planning.

METHOD

PERSONNEL CONTACTED DURING THE STUDY

Since very little published documentation exists which describes the use of training devices in proficiency assessment, an informal telephone survey was conducted. Personnel were contacted who are involved with the use of training devices at United Airlines, the Federal Aviation Administration (FAA), U. S. Air Force (USAF) and U. S. Navy (USN). A list of these personnel and their specific organizations is contained in Appendix A. Additional information concerning the U. S. Coast Guard and USAF Tactical Air Command was obtained through published documents (References 38, 51).

Surveyed personnel were asked about their knowledge of the use of training devices for proficiency assessment within their branch of service, commercial organization, or agency. Questions included inquiries about methods of constructing tests using training devices and how decisions were made to assign some tasks to operational systems and others to training devices. They were also asked about the nature of performance measurement techniques used in the training devices. Finally, they were asked how tests were used and if their organization had studied the validity or utility of using the tests for those purposes.

Personnel were also contacted within selected Army organizations in order to obtain information concerning the current use of training devices in Army testing programs. These personnel were at the Individual Training Evaluation Directorate (ITED) at the Army Training Support Center (ATSC) and also at the Soldier's Manual/Skill Qualification Test Branch at the U. S. Army Infantry School.

REVIEW OF THE LITERATURE

In order to identify research pertaining to the use of training devices in proficiency testing, searches were made, both manually and using automated services, of the general psychological and training literature. Since few articles were found that were specifically directed to this narrow topic, a broader manual search was made covering topics such as: training device design, instructional systems development, training device training effectiveness evaluations, the use of performance tests in job proficiency assessment, complex performance research, automated performance measurement, and utility of measurement.

Additionally, Army publications (*e.g.*, field manuals, training circulars, regulations, pamphlets, etc.) were searched to provide information about Army training and evaluation programs and the processes used to determine unit readiness.

FINDINGS OF THE LITERATURE REVIEW AND SURVEY

USE OF TRAINING DEVICES IN EVALUATION PROGRAMS

Background. The application of the systems approach to training technology has produced extensive innovations in current military and industrial training and evaluation programs (Reference 19). Two salient aspects of this approach provide a background for understanding the current use of training devices: (a) all training goals are stated in objective, measurable terms that are directly related to the performance of job tasks; and (b) there is an increased emphasis on evaluation of performance because of the importance of quality control in training programs and feedback in the systems approach (Reference 6).

Performanced-based, job-relevant, training objectives requires equipment for performance-oriented training. Training devices are used increasingly to satisfy this need because they are usually cheaper to acquire, safer to operate, and less demanding to maintain. Training devices require less energy, less personnel, facilities, and time (References 3, 35, 53, 75). Therefore, training device use for performance-oriented training is increasing.

More directly related to the present study, are the proficiency assessment and certification programs that use training devices as part of their mix of measurement tools. There is a long history of simulation in proficiency measurement (References 18, 27, 29). In addition to training management, proficiency data generated in training devices have also been used for many aspects of personnel management, including selection, certification, and promotion (References 15, 35). It has also been suggested that performance measurements obtained with training devices can play a significant role in research on complex human performance (References 1, 2, 30, 57), in investigations of tactics and doctrine (Reference 35), and in evaluations of readiness (References 18, 38, 77).

The remainder of this section provides an overview of the use of training devices in proficiency assessment programs by the commercial airlines, the FAA, the USAF, the USN, and the U. S. Coast Guard. The overview emphasizes programs that provide performance information describing the proficiency of personnel in operational or line organizations. These personnel receive periodic examinations and certification of their skills relevant to their current jobs or jobs higher in their career ladder.

Commercial Airlines and the FAA. Commercial airlines and the FAA use simulators in their training and evaluation programs more extensively than the military (Reference 11). The FAA has approved the use of simulators for proficiency assessment and certification for a number of commercial airlines (*e.g.*, American, Braniff, Delta, Northwest Orient, TWA, and United) and for a number of different aircraft (*e.g.*, B-727, B-747, L-1011, and DC-8).

The airlines began using simulators for transition training in 1967 and their use for proficiency assessment developed as an extension of these programs (Reference 35). Now, not only do pilots receive part of their rating checks in the simulator at the end of transition training, but also return to the simulators, which are located at central training facilities, for part of their periodic certification programs.

There has been no published documentation or research on the procedures used to construct tests for these simulators. For example, there is no published documentation of formal procedures for determining the cost-effectiveness of using simulators for testing certain tasks as opposed to using the parent aircraft. According to personnel at the FAA and in commercial airline simulator sections, an informal feasibility analysis is performed using subjective opinions of whether or not a task can be performed and tested in the simulator. Thus, the main criterion for use of the simulator for a task is acceptance by instructors, pilots being tested, and FAA personnel. If pilots do not accept the simulator for a given test, then it will not be implemented. In fact, even after the simulator is established as a flight-checking tool, it is possible for pilots who have failed the simulator flight checks to retake tests in the parent aircraft. This is rarely done, however.

For the most part, tasks tested in the simulators are tasks which are proceduralized (e.g., instrument checks) or dangerous, if not impossible, to assess in the aircraft (e.g., emergency procedures).

The procedures used to construct tests covering these tasks are the same as those used to construct tests for the aircraft. As part of instructional system development, training and evaluation objectives are developed. These evaluation objectives are used to determine test tasks, conditions, and standards. Performance is then measured using the same procedures that are used in the aircraft (e.g., checklists). Automated performance measurement capability is available for a few simulators, but lack of acceptance by the pilots has prohibited its use.

Airlines have usually not made public data on the validity of tests conducted in simulators. Williges, Roscoe, and Williges (Reference 75) reported results of studies conducted by Trans World Airlines (Reference 66) and American Airlines (Reference 3) which they said demonstrated that proficiency checks conducted in simulators "accurately predicted performance in the corresponding aircraft". Williges and her associates suggested that the FAA allow increased use of simulators for proficiency assessment for commercial airlines.

Given the lack of extensive research demonstrating the effectiveness of using simulators for proficiency assessment, or even for training (Reference 12), personnel (particularly pilots) involved in the use of commercial aviation simulators have developed a strong belief in the need for high physical fidelity in the design of simulators (References 12, 35). Federal Airline Regulations and other FAA documents that deal with FAA approval of simulators, set forth guidelines for assessing the physical fidelity of the simulator to its parent aircraft. Displays, controls and handling characteristics are assessed in terms of their fidelity to their counterparts on the aircraft. For example, FAA Advisory Circular AC 121-14A states that "the rate of change of simulator instrument readings and of control forces should correspond to the rate of change which would occur on the applicable aircraft under actual flight conditions for any given change in forces applied to the controls, in the applied power, or in aircraft configurations" (Reference 49).

All individual simulators are inspected periodically to insure they are still calibrated and meet physical fidelity standards. They are given frequent check rides, just as individual aircraft are, to determine if they are mechanically fit to be "flown" for training or evaluation.

The FAA also has regulatory control over general or private aviation. It has been suggested (References 75, 49) that training devices could be used for proficiency assessment of private pilots, but currently this use is very limited. For example, FAR 61.57 allows an instrument-rated private pilot to regain his instrument flight certification by means of a flight check, "part or all of which may be conducted in an aircraft simulator or pilot ground trainer equipped for instrument flight and acceptable to the FAA" (Reference 49).

Pilot ground trainers are different from the sophisticated simulators used by the commercial airliners. They are part-task trainers that are not designed to match any specific aircraft. They simulate fewer characteristics and have less fidelity than most aviation simulators. They vary in capabilities and equipment from model to model.

Ontiveros (Reference 48) recently conducted an experiment to determine which capabilities and equipment were needed on pilot ground trainers in order to use them in specific flight tests. Although the specific aspects of this study are discussed later in this report, it is important to note that the pilot ground trainer, with certain equipment and capabilities, was effective as a part-task flight-checking device.

In summary, simulators are used for proficiency assessment and are fully accepted by commercial airlines and the FAA. In general, tests are constructed and administered in these devices as if they were aircraft. In some cases, a check in the simulator is the only test given before a pilot flies under operational conditions. It must be remembered, however, that he flies as part of a crew and that pilots are a sophisticated test population with considerable professional skills.

Air Force, Navy and Coast Guard. The Air Force and Navy mainly use aviation simulators, although there are an increasing number of devices employed in maintenance training and in non-aviation programs (*e.g.*, ships and submarines). Major Navy (*i.e.*, OPNAV 37-10) and Air Force (*i.e.*, AFR 60-1) regulations concerning standards and evaluations (stan/eval) in aircraft allow restricted use of simulators for certain flight checks, but this use is subject to the prerogatives of commanders within major commands, wings, and squadrons (Reference 49). Thus, in this sense, there is little standardized usage.

The surveyed Air Force and Navy personnel emphasized that, as in commercial aviation, simulators in the military (especially in aviation) tend to be costly and sophisticated, reflecting the complexity of the parent systems. Further, interviewees emphasized that each individual simulator facility is very likely to have its own unique program of use in training and evaluation which varies from year to year with changes of commanders and training/evaluation philosophy. For example, the major Air Force Commands - the Tactical Air Command, Strategic Air Command, and Military Airlift Command - all have different regulations and standardization/evaluation (stan/eval) philosophies regarding evaluation and the use of simulators in evaluation. This is complicated by differences within aircraft communities, reflecting their different missions and capabilities (e.g., F-4, F-111) within a command. Furthermore, Naval personnel stated that there were differences in the use of the same simulator for a given aircraft (e.g., S-3) between east and west coast locations of the simulation facilities due to differences in missions and command stan/eval philosophies.

Adding to these reasons for variability, simulators themselves differ in capability because of: (a) complexity of design (e.g., distinctions are made between full mission rehearsal simulators with visual and motion capabilities versus training devices with lower physical fidelity); (b) sophistication and flexibility of the equipment (e.g., older analog systems versus newer digital-based designs); and (c) degradation in simulator performance due to calibration, maintenance, logistics, and personnel problems.

In response to the need to document this diverse use of simulators, as well as the need to organize the extensive research conducted by different agencies within the government and industry, the Air Force Human Resources Laboratory has recently initiated a large scale investigation called the Simulator Training Requirements and Effectiveness Study (STRES).

Several of the products of this project will be of direct use to Army research programs. To wit:

- A detailed review of the use of aviation simulators in evaluation by military and government agencies (as a small part of the total survey goal of documenting simulator use).

- Extensive cost models comparing the use of simulators and operational systems as well as construction of a worth of ownership model.

- Analysis and synthesis of data to identify factors (e.g., simulator fidelity, performance measurement features) which influence simulator effectiveness (e.g., transfer of training, user acceptance and utilization).

Since it was beyond the scope of the present effort to conduct an extensive review, a few selected Navy, Air Force and Coast Guard programs were reviewed to illustrate general trends in the use of simulators for evaluation.

Use in the Antisubmarine Warfare (ASW) Community. Recent Naval regulations have used training devices for the training and evaluation of readiness for operational air and surface ASW platforms. In the aviation community, COMAIRASWINGONE INST-C-3500.2, the Training and Readiness Manual promulgated by the Air Wing One Command, calls for a series of ASW qualification exercises to be completed as part of the squadrons readiness training. The manual describes various scenarios and, although it does not specify which scenarios should be used in the simulators, it permits the simulators to be used for evaluation. The performance of the crew is scored as it is in the aircraft (_i.e._, pass-fail using subjective checklists) although personnel at the facility pointed out that it is easier to control and observe the test.

Successful completion of the scenarios run in both the simulator and aircraft are included for assessment in the squadron's monthly readiness report. Personnel at the squadron said the simulators were accepted because of their high physical fidelity and because the simulator was able to present a wide range of tactical situations which are ususlly prohibitively expensive or impossible to perform in the aircraft.

A different situation exists in the surface ASW community where the 14A2 ASW team trainer is used in readiness improvement programs as specified in COMNAVSURFPACINST-C-3590.1. A review (Reference 17) of this program noted that objective reliable measures were not available for use in the trainer. Furthermore, the tests conducted in the trainer were not and could not be validated because valid criteria of readiness did not exist. Thus, in actuality, the reviewers concluded that the simulator could not be used in determination of fleet readiness although it plays a major role in readiness training.

Use in the Military Airlift Command and Tactical Air Command. The Military Airlift Command has missions similar to the commercial airlines and thus has a simulator program that, according to some Air Force personnel, is one of the most advanced in the Air Force. The program is based on the United Airlines transition training program. The various operational wing commands have simulators that are used for skill maintenance training and, to a varying extent (depending on the aircraft, nature of the simulator, command emphasis, and unit stan/eval philosophy) evaluation.

Flight checks conducted in the simulator are informally constructed through cooperation of the training, stan/eval and operations offices in the wings, and at command headquarters. The major criterion for selection of the simulator is acceptance (of the use of the simulator to measure a given task), cost, and availability of aircraft. The scoring in the simulator is the same as in the aircraft, (_i.e._, subjective checklists).

The Tactical Air Command has missions (_e.g._, air intercept, close support) that are difficult to simulate. Tactical Air Command regulations have a provision for using simulators when aircraft can not be used because of lack of availability, weather, or other operational constraints. The simulator is not to be a primary evaluation tool (Reference 38). It is normally used in a supplementary fashion, providing preliminary checks before checks are conducted in aircraft, as opposed to replacing the use of the aircraft.

Coast Guard. The Coast Guard has its major aviation simulators at its Aviation Training Center in Alabama. All U. S. Coast Guard helicopter pilots annually return to the center for one week of intensive training and evaluation in instrument and emergency procedures, which are conducted completely in the simulator (Reference 51). Upon completion of training, the pilots instrument rating is renewed for another year.

Evaluation procedures in the Coast Guard simulators are similar to those found in other programs. The major use of evaluation is for training management (i.e., identifying weaknesses in individual pilot performance and assessing the training program itself).

Summary Comments on Current Uses of Training Devices in Evaluation Programs. There is a considerable variation of the use of simulators in evaluation and training programs by military and civilian agencies. Moreover, there are few published documentations of this use and there are no formal studies of costs or effectiveness.

Evaluations in simulators are used mainly for the purpose of training management as part of preliminary training programs, transition training programs or recurrent maintenance or readiness improvement training. Uses are also made for personnel management programs to provide data for proficiency certification.

Where simulators are used to test proficiency attainment or maintenance, tests are constructed and administered using the simulator as if it were the parent system. The decision to use the simulator for measurement is based mainly on the acceptance by all involved of the ability of the device to faithfully duplicate whole or part-task situational variances. The tremendous cost of acquiring and operating operational systems has been a factor in the acceptance of simulators for proficiency testing. Other factors cited by personnel included: (a) the unavailability of operational systems; (b) the capability of some simulators to evaluate tasks which cannot be performed in the parent system for reasons of safety (e.g., emergency procedures or the firing of missiles) or security (e.g., the operation of electronic warfare equipment); and, (c) the existence of tasks that cannot be performed in the parent system except in combat (e.g., strategic bombing missions, ASW missions).

EVALUATION IN THE ARMY

This section presents a general overview of the major Army evaluation programs in order to illustrate: (a) the types of tests used and problems with them; (b) the types of organizations, and their resources, that construct and administer tests; and, (c) the types of organizations that use test information and how they use it. It is only within the context of these Army programs and organizations that the effectiveness of tests using training devices and their costs can be determined.

The most well organized evaluation program in the Army is part of the Training and Evaluation System (Reference 61) which is the foundation of the Enlisted Personnel Management System (EPMS) (Reference 39). The first part of the section reviews this program to emphasize the role of performance testing in the Army within the context of a systems approach. The next part discusses the role of performance tests in the Unit Readiness Reporting System. Finally, the use of simulators in Army aviation is considered to illustrate recent changes in simulator use for evaluation.

The Army Training and Evaluation Systems and EPMS. The U. S. Army Training and Doctrine Command (TRADOC) has recently begun the implementation of a comprehensive personnel subsystem program, the EPMS, designed to improve the training, evaluation and management of enlisted personnel. The Army recognizes that the proficiency of its personnel is an important component, along with weapons capability and tactics, in the determination of combat readiness.

The basic building blocks of the EPMS are (Reference 39):

- The specification of critical job tasks and the conditions and standards for their performance. Tasks for individuals are found in Soldier's Manuals for the various Military Occupational Specialties (MOS) within the Army. Tasks for collectives (*e.g.*, platoons, companies, battalions) are found in Army Training and Evaluation Program (ARTEP) publications for various types of units. These publications are developed by personnel within TRADOC schools and centers and within the Army Training Support Center at TRADOC.

- Performance-based, exportable training. The majority of Army training must take place in operational units. In recognition of the problems of decentralizing training from TRADOC schools and institutions (which have staffs and other resources to support their training mission) to operational battalions (which have very limited resources to support training), TRADOC has developed exportable training and evaluation programs to be used by the battalions in the field. These materials include the Soldier's Manuals anr ARTEPs (which provide training and evaluation objectives), field and technical manuals and Training Extension Courses. Training devices, previously concentrated within central training institutions, are now being procured for distribution to field units.

- Skill Qualification Tests (SQTs) and ARTEPs. The SQTs measure individual performance of tasks specified in the Soldier's Manuals. ARTEP evaluations measure collective performance on the tasks specified in the ARTEPs. They are the corner stones of the EPMS system in that they drive the training and personnel management systems.

The core of the EPMS is a Training and Evaluation System. A recent articulation of this system (Reference 61) is reviewed in the next paragraphs.

Operational battalions are evaluated using ARTEPs and SQTs to provide two major outputs (see Figure 1, outputs A and B):

- to the unit (e.g., battalion) trainers to insure that any deficiencies are corrected in the ability of individuals and units to perform their assigned tasks (as defined by Soldier's Manuals and ARTEPs); and
- to the developers of the training and evaluation system to insure the training programs are providing proper and effective training and also to ascertain if the measures themselves are effective.

Figures 2 and 3 present the process by which SQTs and ARTEPs are developed and used. They also identify the agencies involved in this process and the type of information transmitted in the process. For both tests, TRADOC agencies (Training Support Center directorates, proponent schools) develop the basic tests. ARTEPS, which are more complex and less standardized, are modified by local evaluation units selected to evaluate lower level units.

In the conduct of the tests, evaluators are allocated from local or evaluation units. They summarize and transmit the collected data to:

- the chain of command
  -unit trainers
  -unit commanders
  -commanders of parent units
- TRADOC agencies
- personnel management centers.

Tables 1, 2, and 3 provide samples of the type of information which is sent to individual soldiers, companies, and battalions (Reference 23).

Unit trainers and commanders use the information to identify specific individuals and collectives who have weaknesses and to identify which tasks were not performed to the specified criteria. They can thus individualize their training programs to these specific needs. They can also use the information to identify deficiencies in their training programs.

Commanders higher in the chain of command, and their staffs, can use the information to pinpoint weaknesses in their units and provide assistance (resources, command emphasis, etc.) to the units to help correct the problems.

TRADOC agencies use the information to identify deficiencies within the training and evaluation system such as needs for improved measures, improved system definitions, or improved training programs. Proponent agencies may also identify weapon system or tactics deficiencies from the test results.

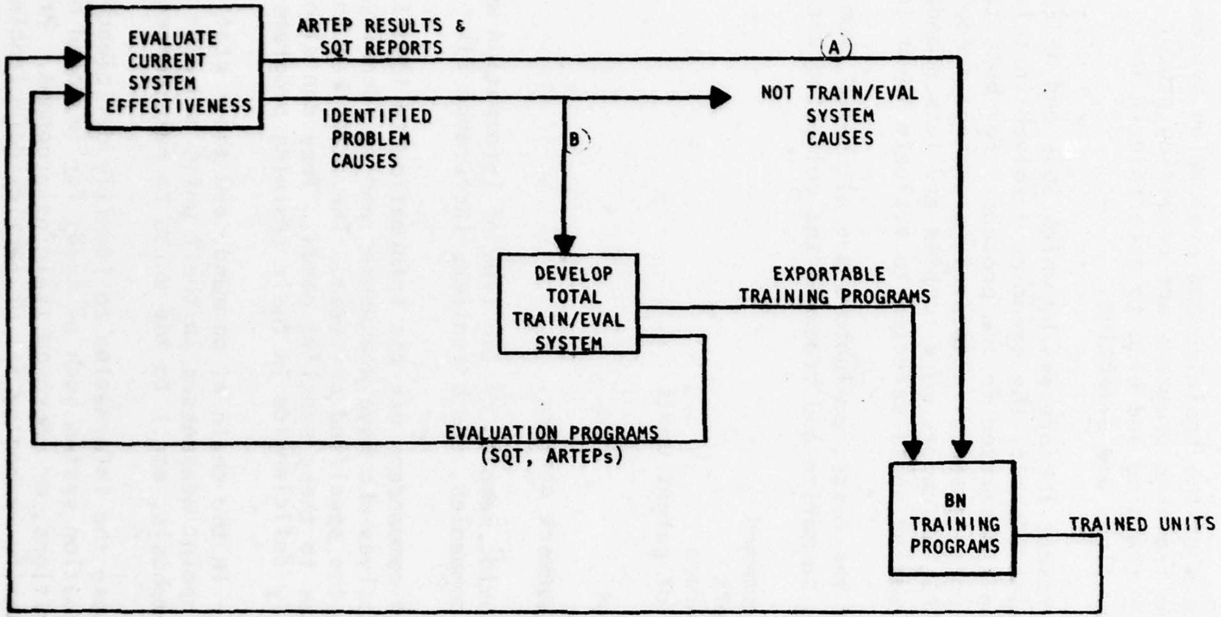
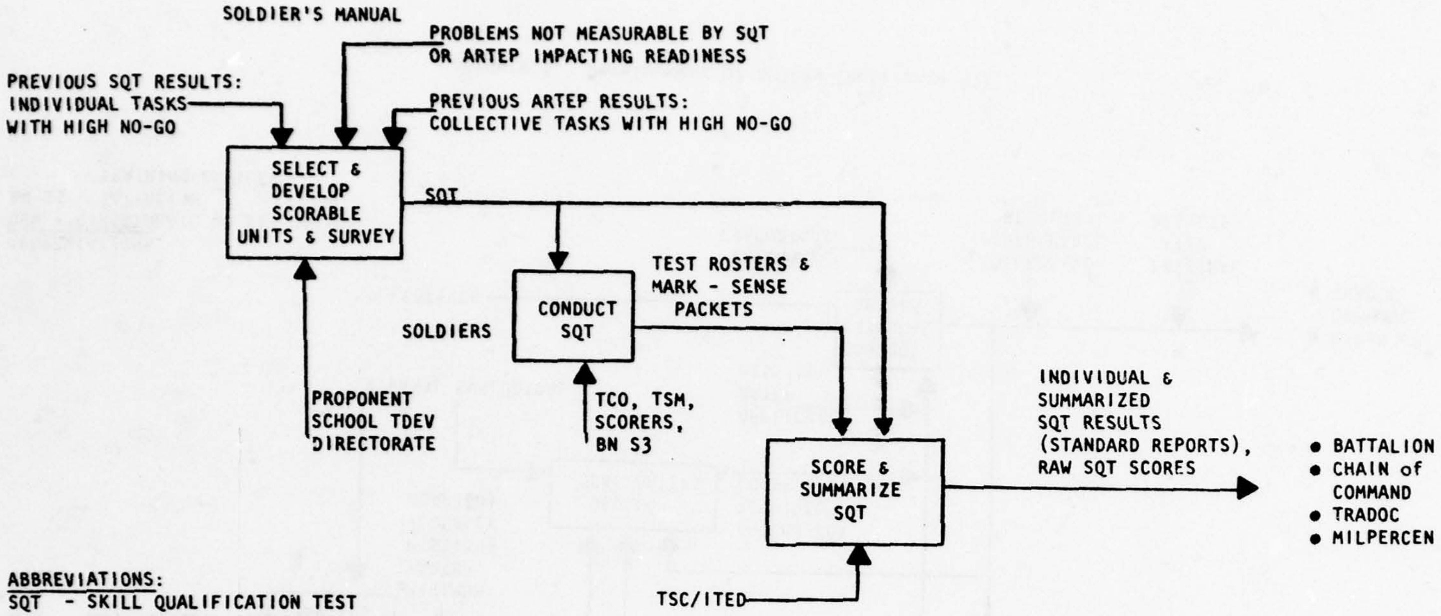

EVALUATE CURRENT SYSTEM EFFECTIVENESS
ARTEP RESULTS & SQT REPORTS
(A)
IDENTIFIED PROBLEM CAUSES
(B)
NOT TRAIN/EVAL SYSTEM CAUSES
DEVELOP TOTAL TRAIN/EVAL SYSTEM
EXPORTABLE TRAINING PROGRAMS
EVALUATION PROGRAMS (SQT, ARTEPs)
BN TRAINING PROGRAMS
TRAINED UNITS

Figure 1. Context of the evaluation subsystem (Reference 61).


SOLDIER'S MANUAL
PROBLEMS NOT MEASURABLE BY SQT OR ARTEP IMPACTING READINESS
PREVIOUS SQT RESULTS: INDIVIDUAL TASKS WITH HIGH NO-GO
PREVIOUS ARTEP RESULTS: COLLECTIVE TASKS WITH HIGH NO-GO
SELECT & DEVELOP SCORABLE UNITS & SURVEY
SQT
PROPONENT SCHOOL TDEV DIRECTORATE
SOLDIERS
CONDUCT SQT
TEST ROSTERS & MARK - SENSE PACKETS
TCO, TSM, SCORERS, BN S3
SCORE & SUMMARIZE SQT
TSC/ITED
INDIVIDUAL & SUMMARIZED SQT RESULTS (STANDARD REPORTS), RAW SQT SCORES
● BATTALION
● CHAIN of COMMAND
● TRADOC
● MILPERCEN

ABBREVIATIONS:

SQT - SKILL QUALIFICATION TEST
TCO - TEST CONTROL OFFICER
TSM - TEST SITE MANAGER
TSC/ITED - INDIVIDUAL TRAINING & EVALUATION DIRECTORATE, TRAINING SUPPORT CENTER
TDEV - TRAINING DEVELOPMENTS

Figure 2. Development of SQTs (Reference 61).

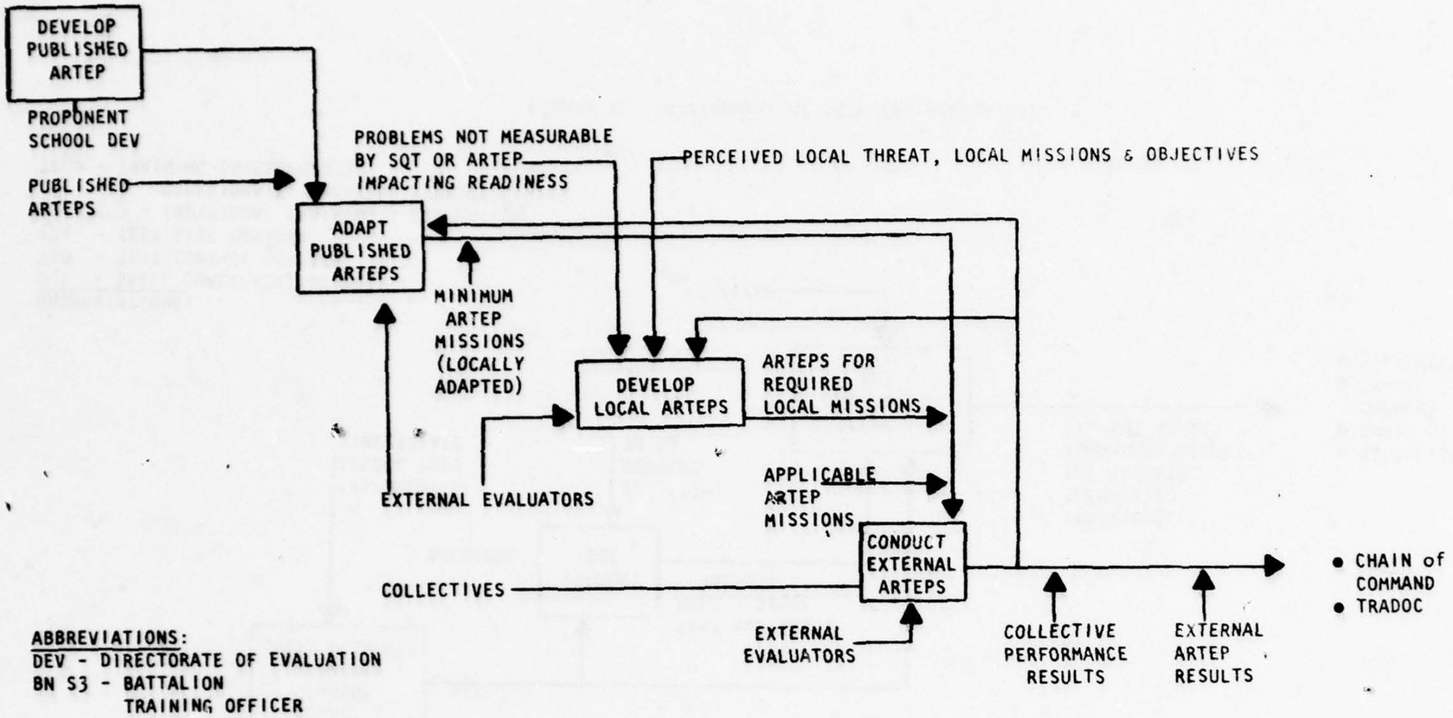
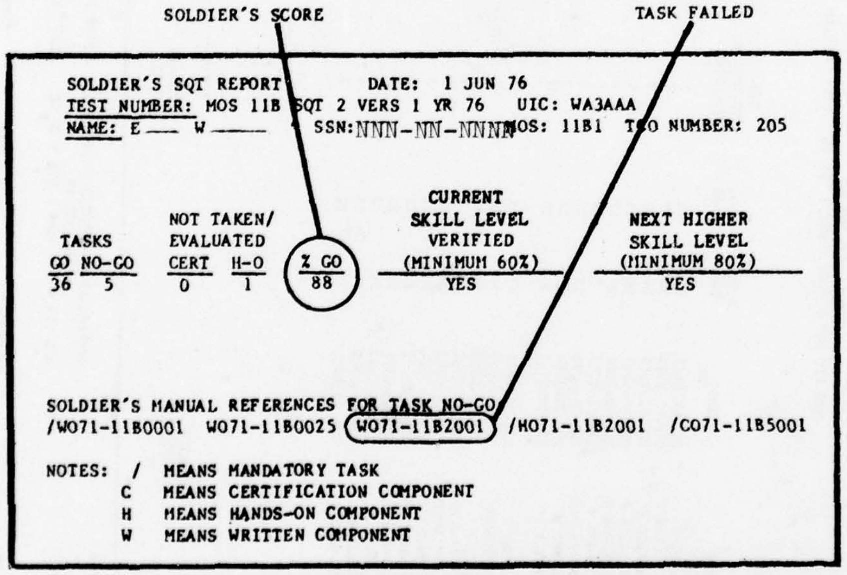

DEVELOP PUBLISHED ARTEP
PROPONENT SCHOOL DEV
PUBLISHED ARTEPS
ADAPT PUBLISHED ARTEPS
PROBLEMS NOT MEASURABLE BY SQT OR ARTEP IMPACTING READINESS
PERCEIVED LOCAL THREAT, LOCAL MISSIONS & OBJECTIVES
MINIMUM ARTEP MISSIONS (LOCALLY ADAPTED)
EXTERNAL EVALUATORS
DEVELOP LOCAL ARTEPS
ARTEPS FOR REQUIRED LOCAL MISSIONS
APPLICABLE ARTEP MISSIONS
COLLECTIVES
CONDUCT EXTERNAL ARTEPS
EXTERNAL EVALUATORS
COLLECTIVE PERFORMANCE RESULTS
EXTERNAL ARTEP RESULTS
● CHAIN of COMMAND
● TRADOC
ABBREVIATIONS:
DEV - DIRECTORATE OF EVALUATION
BN S3 - BATTALION TRAINING OFFICER

Figure 3. Development of ARTEPs (Reference 61).

Table 1

SAMPLE COPY OF AN INDIVIDUAL SOLDIER'S SQT REPORT *


SOLDIER'S SCORE
TASK FAILED

SOLDIER'S SQT REPORT DATE: 1 JUN 76
TEST NUMBER: MOS 11B SQT 2 VERS 1 YR 76 UIC: WA3AAA
NAME: E__ W___ SSN:NNN-NN-NNNN MOS: 11B1 TCO NUMBER: 205

TASKS		NOT TAKEN/ EVALUATED			CURRENT SKILL LEVEL VERIFIED (MINIMUM 60%)	NEXT HIGHER SKILL LEVEL (MINIMUM 80%)
GO	NO-GO	CERT	H-O	% GO		
36	5	0	1	88	YES	YES

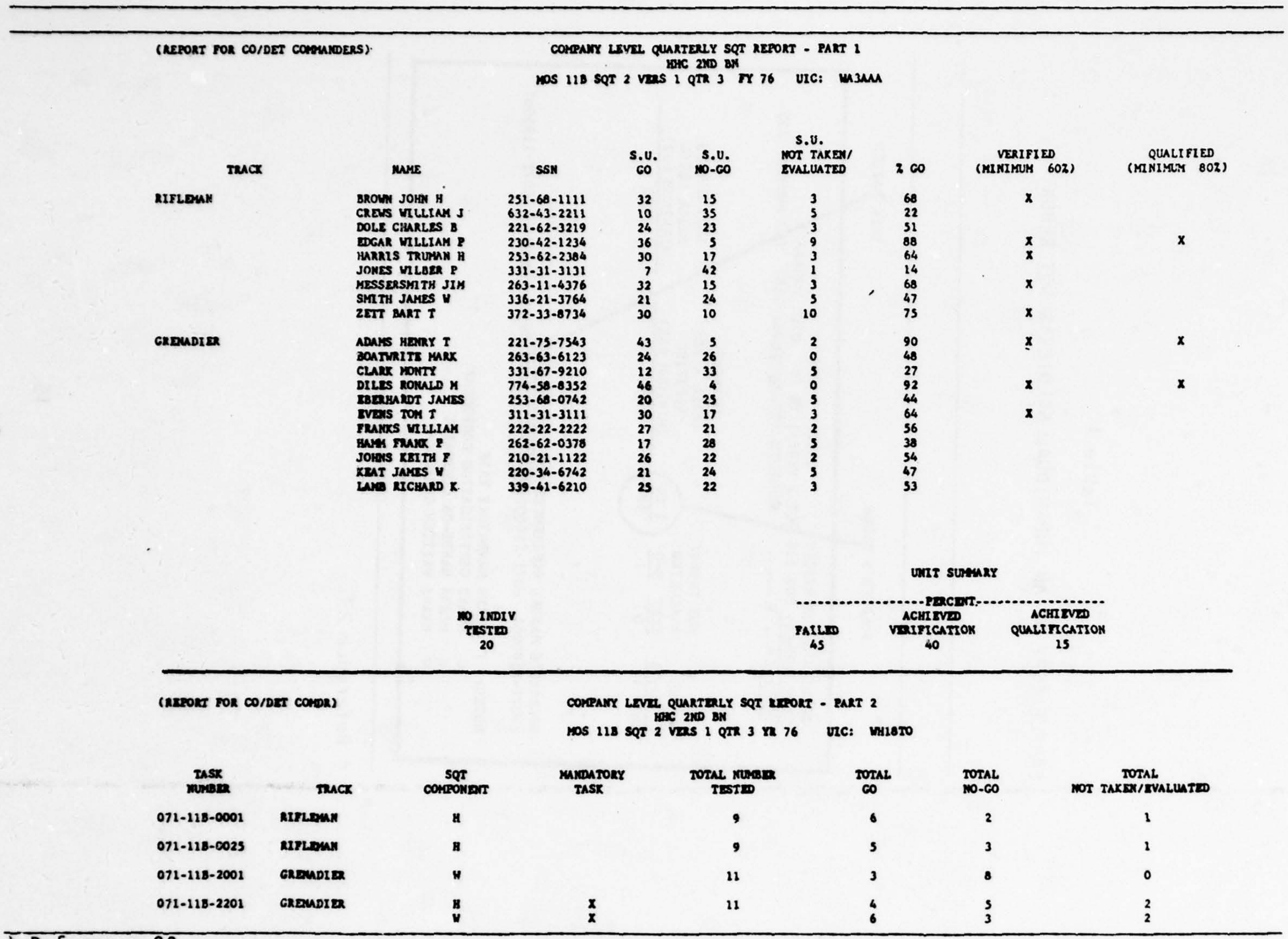
SOLDIER'S MANUAL REFERENCES FOR TASK NO-GO
/WO71-11B0001 WO71-11B0025 WO71-11B2001 /HO71-11B2001 /CO71-11B5001

NOTES: / MEANS MANDATORY TASK
C MEANS CERTIFICATION COMPONENT
H MEANS HANDS-ON COMPONENT
W MEANS WRITTEN COMPONENT

* Reference 23.

Table 2

SAMPLE COPY OF A COMPANY LEVEL SQT REPORT.*



(REPORT FOR CO/DET COMMANDERS)

COMPANY LEVEL QUARTERLY SQT REPORT - PART 1
HHC 2ND BN
MOS 11B SQT 2 VERS 1 QTR 3 FY 76 UIC: WA3AAA

TRACK	NAME	SSN	S.U. GO	S.U. NO-GO	S.U. NOT TAKEN/ EVALUATED	% GO	VERIFIED (MINIMUM 60%)	QUALIFIED (MINIMUM 80%)
RIFLEMAN	BROWN JOHN H	251-68-1111	32	15	3	68	X	
	CREWS WILLIAM J	632-43-2211	10	35	5	22		
	DOLE CHARLES B	221-62-3219	24	23	3	51		
	EDGAR WILLIAM P	230-42-1234	36	5	9	88	X	X
	HARRIS TRUMAN H	253-62-2384	30	17	3	64	X	
	JONES WILBER P	331-31-3131	7	42	1	14		
	MESSERSMITH JIM	263-11-4376	32	15	3	68	X	
	SMITH JAMES W	336-21-3764	21	24	5	47		
	ZETT BART T	372-33-8734	30	10	10	75	X	
GRENADIER	ADAMS HENRY T	221-75-7543	43	5	2	90	X	X
	BOATWRITE MARK	263-63-6123	24	26	0	48		
	CLARK MONTY	331-67-9210	12	33	5	27		
	DILES RONALD M	774-58-8352	46	4	0	92	X	X
	EBERHARDT JAMES	253-68-0742	20	25	5	44		
	EVENS TOM T	311-31-3111	30	17	3	64	X	
	FRANKS WILLIAM	222-22-2222	27	21	2	56		
	HAMM FRANK P	262-62-0378	17	28	5	38		
	JOHNS KEITH F	210-21-1122	26	22	2	54		
	KEAT JAMES W	220-34-6742	21	24	5	47		
	LAMB RICHARD K	339-41-6210	25	22	3	53		

UNIT SUMMARY

NO INDIV TESTED	PERCENT FAILED	PERCENT ACHIEVED VERIFICATION	PERCENT ACHIEVED QUALIFICATION
20	45	40	15

(REPORT FOR CO/DET COMDR)

COMPANY LEVEL QUARTERLY SQT REPORT - PART 2
HHC 2ND BN
MOS 11B SQT 2 VERS 1 QTR 3 YR 76 UIC: WH18TO

TASK NUMBER	TRACK	SQT COMPONENT	MANDATORY TASK	TOTAL NUMBER TESTED	TOTAL GO	TOTAL NO-GO	TOTAL NOT TAKEN/EVALUATED
071-11B-0001	RIFLEMAN	H		9	6	2	1
071-11B-0025	RIFLEMAN	H		9	5	3	1
071-11B-2001	GRENADIER	W		11	3	8	0
071-11B-2201	GRENADIER	H	X	11	4	5	2
		W	X		6	3	2

* Reference 23.

Table 3

SAMPLE COPY OF A BATTALION LEVEL SQT REPORT.*

(REPORT FOR BN/GROUP/BDE COMMANDERS)

BATTALION LEVEL QUARTERLY SQT REPORT - PART 1
1 BN
MOS 11B SQT 2 VERS 1 QTR 3 YR 76 UIC: WH18AA

UNIT		TRACK	TOTAL TESTED	NUMBER VERIFIED (MINIMUM 60%)	PERCENT VERIFIED	NUMBER QUALIFIED (MINIMUM 80%)	PERCENT QUALIFIED
A CO		RIFLEMAN	50	20	40	8	15
		GRENADIER	61	25	41	10	16
		MACHINE GUNNER	20	10	50	3	14
	UNIT TOTAL		131	55	42	21	16
B CO		RIFLEMAN	43	17	39	5	13
		GRENADIER	50	19	38	7	14
		MACHINE GUNNER	47	20	42	7	16
	UNIT TOTAL		140	56	40	19	14
	TEST TOTAL		400	164	41	64	16
	OVERALL TOTAL		500	200	40	75	15

(REPORT FOR BN/GROUP/BDE COMMANDERS)

BATTALION LEVEL QUARTERLY SQT REPORT - PART 2
1 BN
MOS 11B SQT 2 VERS 1 QTR 3 YR 76 UIC: W3YTAA

MANDATORY TASKS MISSED:

TASK NUMBER	COMPONENT	UNIT	PERCENTAGE (MISSED RATE EQUAL TO OR GREATER THAN 25 PERCENT)	TOTAL NUMBER TESTED
071-11B-0015	W	HHC	70	20
		C CO	30	65
		B CO	40	65
		ACO	25	70
071-11B-0027	H	B CO	40	65
		HHC	28	20
		3 CO	28	65
		A CO	25	70

OTHER TASKS MISSED:

TASK NUMBER	COMPONENT	UNIT	PERCENTAGE (MISSED RATE EQUAL TO OR GREATER THAN 40 PERCENT)	TOTAL NUMBER TESTED
071-11B-0001	H	C CO	60	10
		A CO	55	10
		B CO	40	18
		D CO	40	15
071-11B-0002	H	A CO	50	40
		C CO	45	65
		HHC	40	65

* Reference 23.

Personnel management agencies use the SQT results as a partial basis for placement, promotion and other personnel decisions.

The SQT/ARTEP evaluations are considered as "external" tests because they are developed outside of the battalion and their results can be used outside of the unit. Because SQT/ARTEPs can only be conducted at infrequent intervals (e.g., 2 years), the training and evaluation system calls for more frequent "internal" tests to be constructed, administered, and used within the battalion (References 23, 61).

TRADOC PAM 350-X (Reference 23) calls for mini-SQTs to be conducted within the units to aid unit trainers and commanders with their training mission. The results of these tests will eventually be kept in job books that list each individual's proficiency by task.

The Unit Readiness Report. Another major system that guides the evaluation process in operational units is the Unit Readiness Report. Authorizations and procedures for establishing a unit readiness measurement and reporting system are contained in AR 220-1 (Reference 24). The system attempts to provide uniform readiness standards and reporting procedures to insure accurate reporting of combat readiness data to the U. S. Army Forces Command (FORSCOM) commanders.

This section provides a brief overview of the system to provide an additional example of how performance tests are used in operational units. It will also review the severe limitations with these tests and the inherent deficiencies in the reporting system itself.

According to AR 220-1, the unit commander has to employ the readiness reporting system to assess the overall readiness of his unit. This is accomplished by assignment of a overall readiness condition (REDCON) code that best describes his unit's ability to accomplish its assigned mission. Table 4 contains a definition of the combat readiness rating codes.

The overall readiness code is based on ratings, using the same basic code, of the three readiness components people, training status and material. These ratings are based on statistical factors (e.g., personnel strength and equipment inventory relative to the authorized levels) and judgmental factors (e.g., training status, morale, leadership).

It is in the rating of training status that the commander is supposed to weigh various factors (undefined in AR 220-1) and then describe the proficiency of his troops. Col. Irving Heymont (Reference 34) has recently reviewed this process and, in agreement with other experts, found the process extremely deficient.

To begin with, readiness objectives as expressed in the Unit Readiness Reporting System (or elsewhere) are now too broadly defined to enable a commander to judge the readiness of his troops or make decisions on how to allocate resources to improve readiness training

Table 4

COMBAT READINESS RATING CODES[*]

Code	Definition
REDCON 1 (C-1)	Fully ready (C-1). A unit fully capable of performing the mission for which it is organized or designed. Units may be deployed to a combat theater immediately.
REDCON 2 (C-2)	Substantially ready (C-2). A unit has minor deficiencies which limit its capability to accomplish the mission for which it is organized or designed. Units may be deployed to a combat theater immediately.
REDCON 3 (C-3)	Marginally ready (C-3). A unit has major deficiencies of such magnitude as to limit severely its capability to accomplish the mission for which it is organized or designed. Units will require a period of intensive preparation before combat deployment/employment except under conditions of grave emergency.
REDCON 4 (C-4)	Not ready (C-4). A unit not capable of performing the mission for which it is organized or designed. Units will require extensive upgrading prior to combat deployment.

[*] Reference 24.

or improve the current processes of evaluation. Heymont quotes a recent U. S. Army War College study of the Unit Readiness Reporting System which concluded that the measure of training readiness is valueless because of unrealistic assumptions underlying the process and the lack of objective standards for making the estimates. It recommended a consideration of the elimination of the present procedure because it gives a false picture and "those involved in the reporting system know that actual proficiency is far less than reported."

The lack of objective standards for making assessments of proficiency can be traced to the nature of internal training and evaluation programs within operational units. AR 350-1 states that authority and responsibility for training and evaluation be delegated to the lowest command element, usually battalion level, which has the ability to perform the mission. These units have access to exportable training and evaluation material from major TRADOC commands, but must set up their own programs. Their commanders and trainers develop their own goals in achieving specific levels of proficiency and in measuring this proficiency.

This independence has resulted in a large difference in the variety of tests conducted in separate units and diversity in standards of performance for these tests (Reference 54). There are a number of tests conducted that are constructed from different published references (Training Circulars, Soldier's Manuals, weapon qualification courses, etc.). For example, a review (Reference 54) of marksmanship tests found there were a number of tests conducted within and between units, but there was little or no correlation among the various criteria or standards used in the tests. In addition to their inconsistency, the tests were not criterion-referenced or combat-referenced. Studies of antitank gunnery tests (Reference 34, 63) report similar findings.

As a partial solution to these problems, Heymont (Reference 63) recommended that ARTEPs, modified to correct certain deficiencies, be used to objectify and standardize readiness determination procedures. Hayes and Wallis (Reference 33) and others (Reference 5) have discussed the problems of using ARTEPs in this manner. The major deficiencies within ARTEPs are that they are presently not standardized, objective tests. They are not specific in defining tasks and conditions for evaluations. The standards of performance stated in ARTEPs are often inaccurate, too general and vague.

Hayes and Wallis believed it would be impossible to use current ARTEPs to measure "readiness" because it is impossible to define the relation of ARTEPs to combat readiness. It is also difficult to assess the contribution of uncontrolled factors (e.g., turnover, conflicting mission demands during the ARTEP) to test performance. If the unit has extenuating circumstances that inhibit their effective performance on an ARTEP, the ARTEP would be an unfair measure of their readiness.

Another objection to using ARTEPs (and SQTs) for measuring readiness is that the tests are supposed to provide feedback to correct deficiencies within individuals, collectives, and training programs. It is suspected that the attachment of a formal unit evaluation status to these tests will hinder the usefulness of these tests as part of the training program. The emphasis will be on passing tests, rather than discovering and correcting deficiencies.

In summary, the readiness reporting system itself is deficient and the tests used by commanders to make their judgments appear to be deficient in terms of specifying readiness levels. It is interesting to note that commanders do not have to report the method by which they evaluate training REDCON status. In fact, there is no mention of the word "evaluation" as an input to readiness anywhere in AR 220-1.

Army Aviation. Because of its unique status, evaluation using aviation simulators in the Army needs to be separately discussed. The Army is currently procuring simulators to be located near operational helicopter units at Fort Bragg and Fort Riley.

A recent revision of Army regulations gave these units permission to use these devices for annual instrument checks, but this use is subject to the prerogatives of the local commanders and evaluators. These devices and evaluation programs are still too new to be described since the units are just beginning to implement them.

ISSUES TO BE CONSIDERED IN THE EVALUATION OF THE PROFICIENCY ASSESSMENT CAPABILITIES OF ARMY TRAINING DEVICES

The problems of determining how to employ training devices for proficiency testing are a subset of the more global problems of measurement system development, job proficiency testing, and training device design. The literature in these areas contain analyses of a number of generic issues which impact on the specific problems of assessing training device proficiency capabilities.

This section presents a review of some of the issues involved in the evaluation of the proficiency assessment capabilities of training devices. Since it is useful to consider these issues in the context of larger measurement development problems, a general model of suggested decision processes in measurement system development was used as a framework to organize information obtained from diverse areas of the literature. Since the model serves to structure the review, the overview ofthe model in the next section serves as an outline of the remainder of this section as well as a brief summary of general measurement development procedures.

Overview of the Model. Figure 4 presents a model of the decision processes involved in measurement development which is based on models proposed by Muckler (Reference 44) and also by Vreuls and Wooldridge (Reference 70).

The model presents four steps in measurement development: (a) information needs analysis; (b) initial measurement analysis; (c) pilot testing and selection of final measure sets; and (d) measurement system effectiveness testing. .

The information needs analysis is an essential first step in measurement system development. It is the derivation of general statements describing what information is needed by what user for what purposes. For example, is diagnostic information needed, or status information, or systems effectiveness impact information? Is the user a company training manager, a defense system readiness analyst, or a resource manager? Is the purpose to design a training program, to plan training schedules, or to assess operational readiness to meet a threat? This stage provides the basis for utility vs. cost evaluations and drives the measurement system design process.

Initial measurement analysis covers the decisions made after evaluation objectives have been derived and prior to the actual construction of the measurement system(s). The output of the information needs analysis answers the basic question "what is it you want to know", while the output of the initial measurement

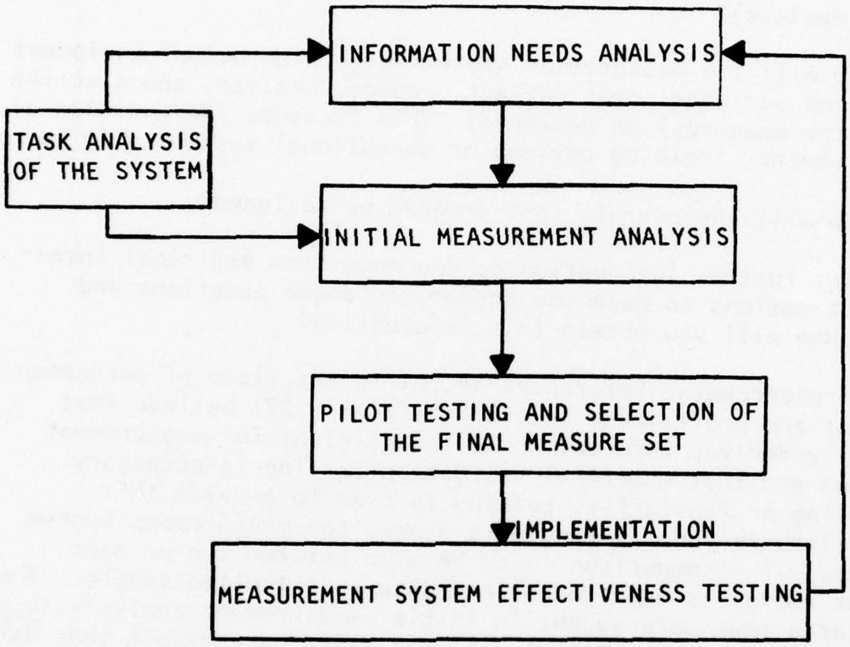

INFORMATION NEEDS ANALYSIS
TASK ANALYSIS OF THE SYSTEM
INITIAL MEASUREMENT ANALYSIS
PILOT TESTING AND SELECTION OF THE FINAL MEASURE SET
IMPLEMENTATION
MEASUREMENT SYSTEM EFFECTIVENESS TESTING

Figure 4. Flow of decision processes in measurement development

analysis answers the questions:

- What is it you want to measure?

- How and within what context will you measure it (e.g., as a part of SQT, ARTEP, or through use of training devices)?

- How will the measurement system (e.g., the actual equipment and personnel that collect, reduce, analyze, and describe the measures) be designed? This includes the decision of whether training devices or operational systems will be used.

- How will the overall test program be designed?

- What further information do you need from empirical investigations to help you answer the above questions and how will you obtain this information?

Many researchers (References 44,70) in the field of performance measurement and proficiency testing (Reference 27) believe that analytically-derived information is insufficient for measurement development and that empirical validation testing is necessary. Pilot testing or feasibility testing is used to provide this empirical information. During this stage, the measurement system (or systems, if a comparison is to be made between two or more assessment tools) is used to collect data on a limited sample. The same criteria that were sought in initial measurement analysis (e.g., utility, validity, reliability) are considered here except that data is collected measuring aspects of the actual use of the measurement system.

The final stage, measurement system effectiveness testing or product testing is employed to determine whether the measures obtained actually were of use to the personnel who use test information (Reference 70). This process is actually a part of larger evaluations which examine the effectiveness of training programs, readiness determination procedures and personnel management programs. The basic question asked here is "were these programs any better because of use of performance tests (in this case, especially in training devices) or improvements in performance measurements"?

It must be noted that this model is presented as a device to structure the following discussion. In reality, these hypothetical stages are merged, omitted or mixed depending on the specific situation in which measures are developed.

Information Needs Analysis. There are at least two issues related to the information needs analysis: (1) As noted earlier, the measurement systems of training devices are often poorly and inadequately designed; a reason for this is the failure to perform an adequate information needs analysis; and (2) the current overall evaluation systems mechanisms may require further development in order to get the information from a training device measurement system to the proper users. With regard to the first issue, a complete information needs analysis is rarely, if ever, performed during training device development. There are numerous decision-makers who use, or could use, information from proficiency tests conducted with training devices in operational units (e.g., units, trainers, commanders concerned with troop and unit readiness, personnel managers, training system managers at higher levels of command and at TRADOC or proponent schools and centers, and researchers concerned with human performance, man-machine system performance, and tactics development). The information needs analysis section will discuss who needs what type of information from performance tests. It is only from within the context of these programs that the objectives of evaluation can be stated, that decisions regarding the design of the training device measurement system and overall evaluation system can be made, and that the ultimate determination of effectiveness can be made.

With regard to the second issue, the problem is that the current evaluation systems, the Army training and evaluation system and the unit readiness program, are usually not designed to accept performance data from training devices. This is primarily because the more complex training devices are relative newcomers to the unit environment and their potential for training and operational readiness assessments have not yet been generally recognized. The impact of this issue is that the information needs analysis should impact on overall evaluation system design as well as on design of the device measurement system.

Initial Measurement Analysis. Given the requirements from the information needs analysis, the next step is to decide:

- What is it you want to measure and how can you best measure it?
- How will the training device measurement system be constructed?
- How will the design and use of the device measurement system be integrated with the design of the total evaluation program?

This stage of analysis is relevant to the present investigation in that the definition of the possible measurement sets drive the selection and design of assessment tools. The tools can include at least SQTs, mini-SQTs, ARTEPs and unit training devices. The question is, What needed information can be provided by each of these tools in a cost effective way? The analyses in this stage are extremely important in the design of performance measurement packages for training devices (References 4, 20, 60).

Often, if this stage is underemphasized, training devices are poorly designed for performance measurement. For example, they may have little or no capability to measure performance because incorrect assumptions were made during initial measurement analysis concerning the ease with which measures could be obtained (Reference 4). Conversely, designers may utilize the "baseball statisticians" approach in which automated instrumentation provides an abundance of information that cannot be cataloged, stored, retrieved, or used (Reference 25).

It must be noted that the following discussion of issues related to initial measurement analysis is based on a systematic approach to measurement development. In operational situations, these procedures may be merged, omitted or underemphasized. However, the formal procedures provide a useful structure to organize and conceptualize the decision and procedures which are inherent in measurement development. It is felt that informal, unsystematic procedures are simply degraded modifications of the formal procedures and that the following discussion is useful regardless of the formality of the employed procedures.

<u>What to Measure and How to Measure It</u>. Major texts on performance testing (References 27, 29, 50, 64, 69) all emphasize that one of the weakest links in the construction of performance tests is the a priori definition of behavior to be measured. The Instructional Systems Development Model, (ISD), commonly used throughout the military provides a framework for the definition of measurement requirements (References 4, 6, 10, 20, 32, 60, 65).

The basic procedures of ISD are designed to aid developers of training and evaluation programs focus on important behaviors in a given system. One of the end products of these procedures should be the design of adequate performance measurement capabilities within training devices and other assessment tools (including the operatial system). There are, however, inherent weaknesses in the translation of training and evaluation requirements into design specifications. The reasons for these weaknesses include:

- The requirements are not specified in terms which can be correlated with engineering options in device design. The device designers find it difficult to formulate these requirements into design criteria which can be traded-off against other criteria (*e.g.*, cost, space).

- The procedures are usually done too late to impact on the design process or the results simply never reach the designers.

- Training specialists have insufficient data bases and technologies to aid them in producing specifications. For example, Blanchard (Reference 9) recently conducted a survey to determine the perceived utility of human performance data sources and concluded they were only of limited usefulness to personnel attempting to measure complex, operational performance.

There have been several efforts (References 10, 20, 32, 58, 65) to correct these problems through the development of supplemental procedures specifically configured to aid in training device design (including the development of performance measurement packages).

The human performance measurement literature, both basic (References 2, 14, 16) and applied (References 70, 76) notes the problems of analytically selecting measure sets. The "criterion problem" isperhaps the most fundamental and difficult problem in measurement. In his review of a conference devoted to discussing research of complex performance, Chiles (Reference 14) stated that the problem was still not solved and there seemed to be no theoretical or empirical base for its solution. In a more recent review, Alluisi (Reference 2) concluded essentially the same thing. The problems include disagreement among subject matter experts on the appropriate behavior for a given situation and the difficulty of measuring performance in safe situations when the actual performance of interest takes place in dangerous environments (*e.g.*, combat). If criteria for performance in operational systems is difficult to specify, then it is also difficult to identify criteria for other assessment tools such as training devices.

Given the limitations just cited, it is still possible to go about the job of deciding what should be measured and how it should be measured. Most of the cited references contain descriptions of how to accomplish the job, albeit imperfectly. For example, Glaser and Klaus (Reference 29) and Sweezy and Pearlstein (Reference 64) give task analytic and content validation procedures for defining measures. Training device design guides (References 58, 59) also contain steps for defining performance measurements. There have also been many measurement analyses actually performed on operational systems, and/or on their associated training devices (References 41, 45, 46, 47, 71, 72, 73, 77).

It is impossible to select, a priori, a measurement analyses procedure that would be appropriate for the present investigation. The decision would depend on the operational system selected, the nature of its associated training device, the existing task analyses available for use, and the experience of the personnel involved in the analysis. Specialists in this area emphasize the subjective nature of these analyses and the requirements for expertise on the part of the personnel applying them (Reference 70).

Further research is needed in the general area of criterion specification particularly for complex, multi-mission operations that typify Army combat activities. In particular, methods are needed to guide the selection of mission essential tasks and the extraction from task analytic data of required performance standards. Given this methodology, the expression of job-relevant proficiency assessment criteria (conditions and standards) can be generated to structure the measurements required from operational equipment or training devices.

Another issue related to the questions of what to measure and how to measure it is the utility of the information obtained and the cost of obtaining it. Reflecting on the link between the initial information needs analysis and the subsequent evaluation of the utility of the information, Edwards (Reference 25) has suggested four basic questions which need to be answered and are largely answerable analytically from the initial analysis:

- Whose utility? The persons or organizations need to be identified who are making decisions and whose utilities need to be maximized.

- Utility for what purpose? What are the decisions that require the information? In most cases, those involve management-level allocation of resources to competing programs, groups or individuals.

- What are the values of the possible outcomes of the decisions? What are the stakes? What happens if allocations are not properly made? Test costs must be compared with valuation of outcomes not just available test budgets.

- What categories of information can actually help the decision-maker? Given the strategy of the decision-maker, the constraints on his decision, and the cost of obtaining the information, does the benefit of certain information outweigh its cost?

The cost of obtaining the information can be expected to vary as a function of the selection and design of the measurement tools (for example, How much in the way of tactics should the device be capable of simulating to provide desired evaluation information?) The problems in developing a cost model are discussed in a later section.

<u>Development of Performance Measurement Capabilities of Training Devices</u>. Once a determination has been made concerning measurement requirements and measurement techniques the next step is to design the performance measurement system.

There are no completely comprehensive procedures or systematic theoretical structures for the design of training devices (Reference 58). Most reviews of design problems (References 35, 53) discuss the salient issues that training device developers confront. Many of these issues also impact on the development of the proficiency assessment capabilities for training devices. Salient issues include: (a) the determination of fidelity requirements for testing; (b) acceptance of devices as test tools; (c) the development of automated or semi-automated instrumentation for performance measurement; (d) determination of device costs; and (e) problems in the design of overall training and evaluation programs which use training devices.

The rest of this section is devoted to the review of these design issues. The emphasis of the review is on the identification of potential variables that will have to be included in future research assessing training device proficiency testing capabilities.

Determination of Fidelity Requirements. To be able to decide on the fidelity of stimuli and responses which are needed in a given test, test developers must analytically compare the expected effects on performance of each deviation from high fidelity - as done in SQT development (Reference 69) - based on their own experience and knowledge in existing data bases.

The process of analytically determining fidelity requirements is part of the task-analytical procedures discussed in an earlier section. As with the rest of the task-analytical procedures, there is an insufficient data base on which to make these judgments. Given the need to determine fidelity requirements for the use of devices for proficiency testing, there are two areas of research that may provide guidance: research evaluating fidelity requirements for test design, and those investigations concerned with fidelity requirements necessary for training device training effectiveness.

Reviewers (References 18, 27) of the literature on fidelity requirements for test design have noted that little research has been done on this topic and thus the literature provides little or no guidance for specific design requirements.

The literature assessing fidelity requirements for training devices for training purposes is farmore extensive (References 8, 68). Many reviewers (References 35, 52, 75), however, have concluded that major questions concerning fidelity requirements are still unanswered. For example, most of this research has concentrated on fidelity requirements for motion, extra cockpit visual cues and cockpit configurations in aviation simulation. These are all major sources of expense in training device procurement and it would be beneficial to reduce their fidelity if high fidelity is not needed. Unfortunately, the literature contains little specific guidance for designers. This research is plagued with problems including (Reference 75):

- lack of generalizability from oversimplified laboratory experiments;

- inadequate measurement techniques to assess performance in the training device (an interesting point relative to the present report); and

- an abundance of unresearched, uncontrolled factors which confound results.

Major research efforts are underway to develop data bases to clarify fidelity requirements necessary for training effectiveness. These include programs by the Air Force Human Resources Laboratory which will in part attempt to define fidelity requirements by starting with high fidelity and gradually decreasing fidelity along several stimulus and response dimensions while assessing training effectiveness.

Crawford and Beck (Reference 18) suggest using this systematic approach for determining fidelity requirements for proficiency testing. Performance on a device could be assessed at various levels of device fidelity (varying on several different stimulus and response dimensions) in terms of its correlation with performance on the parent system.

In summary, it is highly probable that fidelity requirements will need to be assessed in an experiment on the use of training devices for proficiency testing. Again, the literature provides little guidance for determining which stimulus and response dimensions are critical to simulate (which is why operational personnel and training device design engineers opt for high physical fidelity in all dimensions that are technically possible to obtain).

<u>Possible Incompatibility of Fidelity Requirements for Training and Proficiency Assessment</u>. Another problem in the determination of fidelity requirements is the possible incompatibility of fidelity requirements for training and proficiency assessment. The total value of a given training device design must be considered as a sum of its value for different uses: training, proficiency assessment, selection testing, and research (Reference 35). It is possible, however, that fidelity requirements may have some degree of incompatibility for these different uses. Gagne (Reference 28) who was one of the first researchers to suggest the use of training devices as assessment tools, warned of possible conflict between fidelity and other design requirements for training devices versus those requirements for job proficiency assessment tools.

There are many device design features which exist to optimize its training functions (e.g., simplification of tasks, part-task trainers, feedback to guide students). These features may result in training devices which optimize the learning process for naive students, but are ineffective test tools for assessing experienced job incumbents.

Many researchers (References 35, 53) believe that a possible result of the accumulation of training device research will be a greater emphasis on part-task and lower-fidelity trainers that provide more cost-effectiveness means of facilitating learning than high fidelity simulators. Again, it is possible, that these devices may not prove useful for proficiency assessment.

In summary, fidelity and other design requirements for a given device are a function of tradeoffs between requirements for its different uses. It is possible for device to be cost-effective for training purposes and valueless for proficiency assessment. Conversely, high physical fidelity, which may be necessary for proficiency assessment, may reduce training effectiveness.

Fidelity and Acceptance. As indicated by the review of the use of training devices for proficiency assessment by other agencies, a major determinant is acceptance of a given device by examiners and examinees. One of the major criteria determining this acceptance is their perception of the physical fidelity of the device to the parent system.

There are several reasons for user emphasis on high fidelity. In the absence of research demonstrating the effectiveness of lower-fidelity devices for assessing proficiency, users and design engineers opt for maximum fidelity as a hedge against uncertainty. Fidelity of simulation can also serve as a motivational variable (Reference 35). Since examinees respond to tests in terms of their perceived fairness (Reference 6), it is possible that high fidelity helps to convince the examinee that a test on a device is fair. Another reason is that physical fidelity requirements can be specified in the same engineering terms that are used to specify operational system specifications. The FAA regulations and military (USAF and USN) specifications governing the approval of simulators for proficiency assessment give very detailed procedures for assessing fidelity. There are currently no routinely-applied procedures for specifying any other type of requirements. Thus, fidelity is used because it can be used.

Mackie (Reference 43) has studied factors, including fidelity, that influence acceptance of training devices in terms of their use for training. He showed that acceptance was a function of a complex interaction of fidelity of simulation and many other variables, including characteristics of users and characteristics of instruction. Similar work needs to be pursued within the context of acceptance of devices for use in proficiency assessment.

Acceptance can be viewed as a determinant of the utility of test information. The use of test information depends on the value placed on that information and that value is, in part, a function of the acceptance of the test vehicle (Reference 56). It may be that user ratings of acceptance of a use of a given device for a selected test will be a salient, if not dominant, indicant of test effectiveness.

It is interesting to note that Mackie (Reference 43) and others (Reference 35) believe that acceptance of a given device can be modified through proper program management and utilization of techniques to encourage acceptance of innovation (*e.g.*, training device advocates). Additionally, personnel, involved in the use of Naval aviation ASW trainers, attributed the acceptance of that device to greater fidelity than the operational system when it came to simulating certain complex tactical scenarios.

Instrumentation for Data Collection. Once decisions have been made about the fidelity of a training device, the next major problem is constructing the instrumentation and procedures for collecting, reducing, analyzing, and displaying data. The following section discusses factors which need to be considered in the design of instrumentation and procedures for the collection of data. It emphasizes the relative advantages and disadvantages of developing instrumentation and procedures for use in training devices, versus their use in operational systems. Topics covered include: (a) sources of unreliability in data collection, (b) displays and other features that can be used to aid in data collection, and (c) engineering problems in the design of instrumentation.

There are several major sources of unreliability that can be attributed to deficiencies in data collection procedures. Variability in observer judgments complicates using manual data collection procedures while variability in test equipment can confound automated data collection.

There has been extensive work on improving the reliability of observer ratings (Reference 51) and it is possible, given certain conditions that these ratings can be objective, reliable, and valid (Reference 70). Reliable judgments can be obtained if: (a) observers have no distraction during data collection due to concern for personnel safety (given dangerous equipment) or other work loads (e.g., operating equipment); (b) all observers have agreed on standards and conditions for performance and have objective check lists to structure and record behavior; (c) it is physically possible for them to observe the behavior; and (d) the observers are well-trained and know how to rate the full range of possible behaviors.

In many operational conditions, especially in the military, these requirements can not be satisfied in operational systems or in training devices (References 70, 75). Often objective checklists (if they are developed) that work well in research are not as effective in routine use because of the lack of training for the examiners. Additionally, observers frequently do not have access to behavior. For example, the training devices for the Dragon antitank missile has no external monitoring capability for examiners to use (Reference 62). In operational systems, there is frequently no location to safely position an observer during operation of the equipment.

There are many other factors that contribute to the unreliability of observers. The observers may vary in motivation, experience and attitudes. There are also several other well-known biases in raters (e.g., halo effect) which require consideration during the development of data collection procedures (Reference 64).

In conclusion of these points, the degree of reliability of data collection can limit the usefulness of tests conducted in training devices or operational systems. Observers, who are used at present in both operational systems and training devices, have limitations in their ability to reliably perform their collective tasks. Work is progressing on the development of automated and semi-atuomated instrumentation for data collection, reduction, analysis and display in both operational systems (References 41, 70, 73) and training devices. Training devices seem to have some distinct advantages over operational systems in terms of: (a) the relative features available to aid examiners during testing and (b) the ease of physically attaching instrumentation to devices.

Training researchers (References 4, 36) have suggested that the instructor's tasks be analyzed so that special functions can be designed to aid him in his training duties. Crawford and Beck (Reference 18) have suggested that many of these special instructional features may be of use for proficiency assessment. For example, automated performance monitoring, originally developed for training and research functions (Reference 41) is obviously a benefit for proficiency assessment.

Other features include: automated sequencing of tasks of scenario construction (Reference 52), knowledge-based computer systems (Reference 18), and computer-based instructional systems (Reference 18). These features reduce the examiners work load by automatically setting up problems, individualizing test procedures, re-initializing system operations, and collecting, reducing, analyzing and/or displaying data.

Whereas many old test equipment and training devices had simple repeater instruments for examiner/instructors to use, new display techniques have increased the instructor's ability to analyze performance in real-time or non-real-time (e.g., hold for later reduction and analysis). In addition to timers, counters, and X-Y plotters, many devices have sophisticated displays and hard copy printouts of analyzed data (Reference 52).

These new features are not without problems. For example, personnel connected with certain aviation simulators in Naval and Air Force units are unable to use automated performance measurement capabilities or display capabilities because:

- no one understood how the numbers were generated or what they meant;
- the volume of hard copy output associated with each student's performance was excessive (Reference 55); or
- instructors simply did not know the capabilities of the device.

Charles (Reference 13) documented additional problems with displays and other features designed to aid instructors. The instructors' consoles were poorly designed from a human factors viewpoint and often presented too much information for the instructor to use.

Automation or semi-automation of data collection may still not solve the reliability problem. Automatic systems require special maintenance programs and operating procedures to insure calibration and correct operation of the equipment (Reference 70). A common reason for lack of use of performance test equipment is that it is misaligned, needs other forms of maintenance, or simply does not work (Reference 4).

If the system is to be attached to operational systems performing in operational environments, then the designer must be concerned with weight, size, packaging, power, heat, vibration and noise requirements (Reference 70). There have been several excellent studies (References 45, 46, 47) that discuss the numerous problems of instrumenting operational systems.

Training devices usually provide better possibilities for attaching measurement systems, but most have not been designed to include measurement systems (References 4, 70). Problems include lack of adequate documentation of the training device's mechanical and electrical processes, timing problems, A/D or D/A conversion discontinuities and limited resources to handle additional computational demands. In summary, it is difficult to fit measurement systems to operational equipment and it is also difficult to retrofit measurement systems to training systems which did not have them originally.

Cost Models. Cost is a salient factor in the construction of tests and also in design of training devices. However, there is a paucity of published documentation on cost models for these decisions.

The only formal attempt to integrate costs of evaluation programs was a part of a comparison between various options for implementing Reserve Component ARTEPs (Reference 5). Cost data was collected for the following elements:

- Personnel required for evaluation,
- Travel required,
- Per Diem,
- Petroleum, Oil, and Lubricants (POL),
- Maintenance (repair parts), and
- Ammunition.

These data were reported for:

- Planning evaluation headquarters,
- Evaluator/controller group,
- Support personnel such as range personnel,
- Aggressor personnel,
- Evaluated unit, and
- Attached and supporting units.

Although this list is not long, it begins to illustrate the complexity of constructing a cost model for evaluation programs such as ARTEPs. In terms of sheer mass of data, summarization of the cost information for this study filled a 150 page appendix.

Cost models for training devices (Reference 37) or instructional delivery systems (Reference 10) provide costs for device ownership and indirectly, costs that would be incurred during use of devices for evaluation. Appendix B presents an example of candidate cost elements for both simulators and aircraft which have been proposed for use in models to compare relative training costs.

There are only a few cost models for training devices because of the difficulty of collecting cost data and constructing models. For example, it is difficult to get various elements of cost from discrete sources. Acquisition costs, maintenance costs, operating costs, and other elements are maintained by several different agencies. Accounting systems within each of these agencies differ and it may be difficult to isolate and abstract needed cost information.

Given availability of cost data, cost models are not easy to construct. For example, variance in assumptions made concerning cost tactics, down time, or availability of equipment and personnel can drastically alter obtained cost figures. This problem is further complicated by the problem of non-dollar costs.

Jolly and Caro (Reference 37) stated that non-dollar costs (e.g., safety), may be more important in the determination of the relative merits of training devices than dollar costs. Consideration of diversion of limited unit resources (e.g., time, personnel, facilities, equipment) for the construction, administration, and use of tests is also of great importance. For example, the time necessary for preplanning for performance tests may divert unit personnel from their normal activities and disrupt unit training or operational missions.

Cost models of training devices are an important developmental issue in any future attempt to assess the utility of training devices in proficiency assessment. A coequal concern is the development of cost models for information systems of which the training devices or simulators may become parts. The literature reviewed revealed no such models. If it can be assumed that the training device used for proficiency assessment is a direct substitute for operational equipment and therefore has no incremental or decremental impact on the information system in which it is embedded, the lack of such a cost model may not be important. However, because training devices have the capability of providing different evaluation data and thus altering the existing information system, the assumption of no impact must be questioned, at least initially. Cost models for information systems must thus be defined at least to the extent that basic variables are identified so that the potential for impact can be assessed.

Program Design. Angell (Reference 4) in his survey of the use of performance measurement in training devices found that performance tests were often not used or not used effectively because of poor program design. For example, performance testing is obviously dependent on the continued accurate performance of test equipment. Often maintenance and logistics planning which should be part of test program planning is underemphasized and equipment cannot be used because of lack of proper calibration, spare parts, or other maintenance problems. Training devices at operational units frequently suffer from neglect because adequately trained maintenance personnel and spare parts are not available in field locations.

Another aspect of total program planning is selection, training, and management of examiners. The literature on problems with instructors in training devices provides insight into possible problems in the use of training devices for proficiency testing.

Instructors in some programs are poorly trained in the use of the training devices and also in training and evaluation procedures (Reference 4, 13). Additionally, instructors are frequently technicians who do not have proficiency in the operational system (especially in aviation) and thus are not respected by examinees.

If instructors are drawn from senior personnel, it can also create problems in that often senior personnel see working with training devices as bad for their careers or as just simply boring. Use of a senior personnel can also disrupt unit activities.

There are many other aspects of program design (e.g., administrative, coordination among organizations, scheduling of tests, frequency of tests, and command emphasis) that impact on the effectiveness of a given performance test program. There will be many differences in the implementation of test programs using training devices or those using operational systems. There is, at present, no information on the impact of these differences on the relative effectiveness of different measure sets.

Pilot Testing. Once decisions have been made concerning what to measure and an alternative system has been designed to collect performance data, the next step is to collect data. This data can be used to aid in selecting the final measures and measurement systems that will be used in the performance test (or tests).

Only a few studies exist that directly address the problem of determining what empirical data to collect to aid in the decision of whether or not to use a given training device as a testing vehicle. This section will review the general procedures suggested in these studies. Additionally, the proficiency test literature and performance measurement research literature were reviewed and selected methodologies prevalent in this literature will be presented here to provide additional insight into the problem.

In an early study, Besnard and Briggs (Reference 7) assessed the usefulness of using a maintenance training device for proficiency testing. He employed a between-groups design which varied the difficulty of parallel tasks to be performed in the training device by one group and in the operational system by another. The errors made on the device were similar in number and kind to those made on the operational equipment. The authors interpreted this result as "suggesting" that the simulator , ɔficiency measurements are representative of those obtained on the operational system. They also pointed to a significant savings in test time in the training device because of the omission of unnecessary steps. These steps were impossible to omit in the operational system. The device also had the unique ability to automatically reinitialize its circuitry after an examinee had made a mistake.

A more complex study (Reference 42) compared performance of experienced pilots in an aviation simulator to performance of the same pilots in an aircraft. All subjects initially received two flight checks in the simulator, then flew a similar check in the aircraft. Performance was scored using an objective flight checklist which had high demonstrated inter-observer reliability. The results of this study showed that performance in the simulator predicted performance in the aircraft quite well. All correlations were significant (the highest was +.724). The findings also showed that the presence of simulator motion (which was varied in the experiment) resulted in a greater correlation between simulator and aircraft performance than when there was no simulator motion.

Another aviation study (Reference 49), partially reviewed earlier in this report, compared pilot flight check performance in a pilot ground trainer (a part-task training device) with subsequent performance of the same pilots on a similar check in an aircraft. A comparative analysis of variance between performance scores (as measured by objective flight checklists) in the trainer and in the aircraft did not reveal any significant differences. A task by task subjective comparison of performance was used to identify specific equipment and trainer capabilities that had to be modified in order to improve its test capabilities.

One study (Reference 77) was designed, but not executed, to assess the use of a Navy carrier air traffic control center (CATCC) simulator to evaluate performance. It proposed one of the most complex experimental designs found in the literature. Personnel were to be evaluated first in the CATCC simulator and then assessed using shipboard CATCC. Additionally, there were between group differences, i.e., different teams of personnel performed on different types (classes) of carriers-- although there was only one version of the simulator.

Several dependent variables were identified as candidate proficiency measures and analyses were proposed to aid in the selection of these measures. It was suggested that an analysis of variance be conducted to determine if: (a) between team performance differences occurring in the trainer were consistent with those differences occurring at sea; and (b) there were any differences between the at-sea CATCC performance levels of different carrier classes.

Additional correlation analyses were also recommended to: (a) estimate the significance and comparative strengths of device and carrier performance measure relationships; (b) to evaluate the independence of measures obtained in the device for parsimonious selection of a device measure set; and (c) to evaluate the independence of measures obtained in the carrier for parsimonious selection of onboard measures.

In conclusion, the above studies show the diverse approaches that can be taken to illustrate the comparative value of performance tests in training devices and operational systems. The selection of a given procedure will depend on the experimental situation (e.g., availability of subjects, availability of training devices and operational equipment, availability of measurement techniques for the system, and availability of qualified examiners).

System Cost Effectiveness Testing. To be theoretically complete, measurement systems should receive a continuing empirical evaluation to reveal what is actually gained from using that system (References 44, 70), and at what cost these gains and costs have to be evaluated within the context of the training, personnel, and other programs that use the test information. It would be useful to know how the use of different measurement systems influence personnel utilization, training time, performance quality, readiness determinations, training device or operational system utilization, cost of operations, and other similar indices of program effectiveness.

In other words, does a given performance measurement system provide Army decision-makers in operational units and major Army commands with better data for the cost, relative to other sources of information, to allow them to operate their programs more effectively and efficiently. Other sources of data include not only other types of performance data obtained in different assessment systems, but also other types of information: paper and pencil tests, supervisor ratings, and personnel data concerning factors such as turnover or morale.

Such a cost and information effectiveness evaluation is extremely difficult to perform. There is only one example in the literature (Reference 71) of a measurement system effectiveness test. This study revealed that empirical measurement development resulted in an improved measurement system that produced a 40 percent reduction in time-to-train compared to use o f a system which was derived from analytic means alone. Costs of alternative information system were not treated.

This study accomplished two major purposes: (a) it demonstrated that empirical measurement development techniques (such as those employed in automated performance measurement procedures) can add significantly to the analytical processes used to develop measures; and (b) it illustrated the value of measurement system effectiveness testing as a form of empirical validation of measurement procedures.

One of the reasons it is difficult to perform system effectiveness testing is indices of program effectiveness, if available, are difficult to relate directly to measurement system designs. It is possible, however, to partially solve this problem by having program personnel, who use that test information, subjectively rate the utility of the information with respect to their needs. An example of this procedure can be found in a study (Reference 5) that investigated the cost-effectiveness of alternative procedures for implementing Army Reserve Component unit evaluations using ARTEPs.

A review of background underlying selection of these effectiveness ratings will provide insight into the present problem. Effectiveness was considered to be a function of the extent to which different ARTEP implementation options met the stated objectives for ARTEPs. A preliminary analysis resulted in the identification of three major indicators of effectiveness: adequacy of information, user acceptability, and promotion of readiness gains.

Although the first two indicators proved to be feasible, the third indicator could not be used. It had been assumed that for quantification of readiness gains either training REDCON (from the Unit Readiness Report) or number of weeks to achieve combat proficiency would be used. However, a review of historical data to obtain such indices was abandoned for two major reasons: (a) the four-point readiness scale represented a scale insensitive to all but substantial changes in proficiency; and (b) the study was concerned with the evaluation function of ARTEP (i.e., direct effects due to the measurement system), while readiness gains can be due to many additional factors impinging upon training (e.g, leadership, personnel turbulence, time, and resource constraints).

It was determined that the feedback portion of the ARTEP training cycle (i.e., training objectives, training, evaluation, and training objectives)was the key to assessing ARTEP evaluation effectiveness. It was felt that the extent to which feedback information is timely (current as well as received in time to be used), accurate, and useful (acceptable to users) determined the efficiency of evaluation.

Surveys and questionnaires were constructed to obtain subjective ratings on these three dimensions for each ARTEP implementation option. Respondents for this survey are listed in Table 5.

The ratings were then aggregated, using simple averaging techniques, into option effectiveness indices. These indices were then integrated with cost data to allow cost-effectiveness comparisons to be made for the different implementation options.

In conclusion, this study of ARTEP implementation options demonstrates that measurement system effectiveness testing can be accomplished using indirect ratings. It also demonstrates the difficulty of obtaining objective program effectiveness indices (e.g., readiness reporting) for use in this type of study.

Table 5

RESPONDENTS IN ARTEP STUDY*

Active Army Officer Evaluators (each evaluation)

Reserve Component Evaluators (each evaluation)

Army Readiness Region Officers with ARTEP duties or experience

RC Officer and Enlisted Personnel in each unit undergoing evaluation

Manuever Training Command Staff Officers with ARTEP duties

Branch School Staff Officers with ARTEP duties or experience

CONUSA Staff Officers with ARTEP duties or experience

FORSCOM Staff Officers with ARTEP duties or experience

TRADOC (including USACATB) Officers with ARTEP duties or experience

Study Advisory Group Members

Selected Pentagon Personnel

Selected General Officers

* Reference 5.

DISCUSSION AND RECOMMENDATIONS

A systematic investigation of the performance assessment capabilities of specific Army training devices is obviously very complex. A set of measurement capabilities (e.g., instrumentation, procedures, test scenarios or items, test conditions, standards) or alternative sets of measurement capabilities must be constructed for the training device (and operational system). These capabilities must then be evaluated in terms of their relative cost-effectiveness as compared to measurement capabilities employed in present operational systems or, given a more global view, in terms of their absolute cost-effectiveness with respect to their impact on Army programs (e.g., training, personnel management, readiness determination) and information management.

As reviewed in the preceding section, there are many questions that can be addressed during the construction and evaluation of alternative measurement systems. The purpose of the following section is to discuss possible avenues of research which may be followed in order to answer these questions with respect to the determination of the proficiency assessment capabilities of a selected Army training device.

The following section is divided into five parts. The first part is concerned with the selection of a specific training device for future investigations. The next four parts discuss the construction of measurement capabilities and evaluation of these capabilities. The same measurement system development model that was employed in the findings section (i.e., information needs analysis, initial measurement analysis, pilot testing and system effectiveness testing) is used to structure the section.

SELECTION OF A TRAINING DEVICE FOR INVESTIGATION

The first step in future investigations will obviously be to choose a device that would be useful and practical to investigate. There are three major considerations for the choice:

- the performance measurement capabilities of the device;
- the availability of research and documentation for the device and its parent system which could provide data for the proposed study; and
- the stage of development of the device.

The first consideration, the performance measurement capability of the device, is of prime importance because the true desideratum of the proposed investigation is to assess and, if possible, justify the inclusion of performance measurement capabilities (i.e., the availability of adequate instrumentation and procedures for conducting proficiency tests). As previously reviewed, many devices are developed without proper consideration of measurement needs. It would be desirable to have measurement capabilities on the selected device to allow comparative investigations of the effects of the inclusion or exclusion of any measurement capabilities (such as automated instrumentation to collect and display performance data or previously constructed test scenarios or items) will reduce the amount of work researchers will have to devote to the construction of tests, etc.

A considerable number of current and near-future research programs have the potential of providing data or conceptualizations that could be useful in the present proposed study. Possible sources include research on: the EPMS, task analysis and performance evaluation technology, cost and training effectiveness analysis of various training programs (including analysis of training devices), evaluation and testing of training devices, unit readiness reporting, training management technology, performance tracking systems, embedded training systems, and test and evaluation of the operational systems. For any given training device and its parent system, there will be an associated body of literature generated from these programs, some of which will be devoted specifically to the device and parent system. Obviously, it would be desirable to select devices with associated documentation which provides information on such things as training effectiveness or costs. For example, two of the previously reviewed studies (References 48,77), which dealt directly with the problem of assessing the use of training devices in proficiency testing, based part of their investigations on information derived from training effectiveness studies (References 26,48), that had been previously conducted.

The last consideration - stage of development of the device - is of practical importance. Future research will be concerned only with devices that will be available to operational units in the field as opposed to those located only at training institutions. The stage of development will determine the availability of devices, and availability of empirical data on the use and effectiveness of the device in the operational units. It would be desirable to have the eventual unit-level users of the device be familiar with the device.

Therefore, a review needs to be conducted of all present and near-future Army training devices. This review would serve as the basis for the selection of a device for future research. It could also document: (a) the existing or proposed performance measurement capabilities of these devices; (b) their uses and planned uses in unit-level Army evaluation programs (*e.g.*, SQT and ARTEPs); and (c) their potential for use in evaluation programs. If there is a discrepancy between present use (and planned use) of devices and their potential use, then documentation of this discrepancy would provide evidence of the need for future work. Additional evidence could be derived from delineation of deficiencies in existing or planned performance measurement capabilities of all Army training devices.

INFORMATION NEEDS ANALYSIS

After selecting a specific training device, the next problem is the identification of the possible roles the device could play in Army evaluation programs and tests. The analysis could focus on the information needs as identified (and limited) by present programs (e.g., EPMS) or it could be part of larger investigations of potential applications of performance measurement data (e.g., research methods to aggregate individual performance data and collective performance data into unit readiness indices).

The primary questions to be asked are "who can use the performance test data and how can they use it?" In other words, if the utility of test information (obtained in a training device or any other assessment system) is to be determined, then the questions of "whose utility and utility for what purpose" must be answered.

If there were a coordinated, systematic network of evaluation programs and tests in the Army then determination of test utility would be relatively easy. Given the present status of Army-wide and unit-specific tests, an analysis will have to be conducted to identify the tests in which the training device can be used. This analysis is complicated by the fact that each unit employs a different set of tests and uses the test information in different ways. Thus, the analysis must include a survey of a sample of the units which will use the device. An example of such a survey for rifle marksmanship is found in a recent study by Rosen (Reference 55).

As part of the survey of Army-wide and unit-specific tests, personnel who construct, administer and use (or could use) the test information will also need to be identified. Identification of personnel who construct and administer tests will allow determination of resource constraints and other limitations for these processes. Knowledge of these limitations will aid in decisions concerning measurement system and test program design.

Identification of personnel who use, or could use, test information will provide access to the knowledge of how test information is or can be used. This knowledge can be used as the basis for constructing indices (to be used during system effectiveness testing) which could be employed to assess the effectiveness of test information in Army programs (e.g., reduction in time-to-train). It can also serve as a guide to the development of ratings, (e.g., usefulness, timeliness, accuracy) that indirectly assess measurement system effectiveness. Obviously, the identified users of test information can serve as respondents for these ratings.

Thus, an initial step in the analysis of the proficiency test capabilities of a given Army training device will be the determination of general information needs which have the potential of being satisfied through device use. It may be possible to employ the device as part of existing tests within the EPMS training and evaluation system. It may also be possible to use devices in either existing or new unit-level tests. Users of unit-level test information will have to be identified, as well as their use (or potential use) of the test information.

INITIAL MEASUREMENT ANALYSIS

Given the selection of a specific device (and its parent operational system), the identification of potential users, and the uses of performance test information obtained in the device (and parent system), what to measure and how to design/select measurement systems and test programs have to be determined.

The decisions made during this phase are based only on analytical judgments. The quality of these judgments will be determined by: (a) the availability of data bases; (b) the type of task-analytic technique chosen to structure the judgments; and (c) the knowledge of the personnel involved in making the judgments.

The literature review revealed a paucity of data bases and a plethora of task-analytic techniques. The review also revealed that this stage is the most difficult and least proceduralized step of test construction and measurement system design (and training device design). Hundreds of variables (many of which are specific to each situation) must be analyzed during a selection of behaviors to be measured, identification of conditions for measurement, and the setting of standards of performance.

There are no cookbook answers to the problem of determining what to measure and how to measure it. However, there are systematic approaches that can be used. For example, group decision-making techniques (e.g., Delphi, Nominal Group Technique) can be used to access the diverse requirements for many of the decisions in this stage. There are numerous personnel who can provide input:

- personnel, at major Army commands and in operational units, who use or could use test information;
- personnel, at major Army commands and in operational units, who construct and administer tests;
- personnel who are concerned with the operational system on which performance is to be measured (e.g., system developers, test and evaluation personnel, school and center personnel concerned with tactics of deployment of the system):
- personnel who are involved in design and development of the training device;
- personnel involved in training programs;
- research personnel involved in the evaluation of the proficiency assessment capabilities of training device; and
- research personnel involved in projects to improve readiness determination procedures, performance evaluation systems and other programs which impact on performance test construction and use.

A major deficiency in current procedures to construct test and design training devices (especially with respect to performance measurement capabilities) is the lack of communication links between these personnel. This deficiency has resulted in tests which are not combat-referenced, training devices (and operational systems) which cannot reliably and validly measure performance, performance measurement research which is not suited to current and future Army problems, and other problems which confound the performance measurement problem instead of reducing it.

These personnel can be used formally (using group judgment techniques) or informally (e.g., advisory panels, respondence surveys) in many decisions. In the determination of tasks, conditions, and standards for evaluation, these personnel can provide references for existing task analyses (as part of system development) or evaluation objectives (e.g., those in Soldier's Manuals and ARTEPs).

They can also provide insight into the sufficiency of these exisiting analyses and objectives. For example, the conditions and standards for the various rifle markmanship tests are neither correlated with each other nor are they combat-referenced (Reference 54). Simply using a training device as a substitute in these existing tests is not a solution to rifle markmanship evaluation problems. Completely new task analyses, evaluation objectives and tests may have to be developed.

The combined knowledge of all these personnel will also be needed in selection of assessment systems to be used to measure different tasks. In the simplest case where there is a fixed measurement system designed for both the training device and operational system*, the ultimate problem is to determine the mix of assessment systems which should be used. There are many trade-offs which will have to be made and many criteria to be considered. Thus, an overall structure is needed to organize and integrate the procedures of judging the cost-effectiveness of using a given training device for proficiency testing.

There are two approaches to integrating and organizing criteria for making this judgment:

- determination of the cost-effectiveness (utility of measurement) of alternative assessment tools (e.g., training device versus operational system) in a specific testing program; or
- determination of the cost-effectiveness (worth of ownership) of owning a specific training device given its multiple uses in training and proficiency testing.

* In reality, the decision should be made before measurement systems are designed and thus serve as inputs to the design process.

Worth of ownership models have been proposed to integrate cost and benefit criteria to determine the worth to the training device users of various device instruction features and other device implementation options. Part of the STRES project, which was reviewed earlier, will be an attempt to construct such a model based on an extensive review of the literature and survey of training programs.

The construction of this model is exceedingly complex with difficulties in the quantification of benefits as well as non-dollar costs. To solve some of these problems, the model may take the form of a branching logic (as opposed to an algebraic integration) keyed to critical questions and indexing existing data and knowledge relevant to the question. This same approach may eventually prove to be a useful way to conceptualize the utility of measurement (which is a component of the overall worth of ownership considerations).

One of the original conceptualizations of utility measurement was based on an application of decision theory to personnel testing by Cronbach and Gleser (Reference 2). Simply put, the value of a personnel test lies not just in its psychometric properties (i.e., validity and realiability), but also in its ability to aid the decision-maker who must use the test to make personnel decisions (e.g., selection, placement, classification).

The application of decision theory means that (a) cost considerations, (b) consequences of decisions (i.e, the valuation of outcomes or payoff matrix), (c) strategies for using test information, and (d) constraints on decisions (e.g., quotas in selection programs) are quantified and then algebraically integrated with measures of validity. Since all criteria can be aggregated and indexed with one metric representing utility, the decision-makers' tasks are simplified from the original process of having to consider the criteria separately and unsystematically.

Although the concept of utility of measurement is extremely attractive, Cronbach and Gleser warned of severe limitations in its present state of theoretical development. For example, it is difficult to determine the consequences of many decisions (e.g., what will be the outcome of correct or incorrect judgments) and it is even more difficult to quantitatively valuate those outcomes (e.g., what are the costs of failure to detect lack of markmanship proficiency).

It is also difficult to determine (and quantify as conditional probabilities) the strategies that operational personnel use for relating a given type or category of information to their decision. For example, how does the unit trainer use test information (if he does use it) to determine the proficiency of a given individual? How does a commander use test information to judge the readiness of squads and other collectives in his command?

It is difficult to quantify these component variables and it is impossible to apply Cronbach and Gleser's quantitative model, as it is now formulated, to the needs of the proposed investigation. Criteria of cost and effectiveness, however, will have to be identified and integrated in order for a decision to be made on the use of alternative measurement systems. Group judgment techniques can be used to identify priorities and integrate these criteria to make many decisions. For example, there are many issues which might be considered:

- selection of behavior or tasks to be tested;
- determination of factors (e.g, environmental, situational) affecting performance so that test conditions can be set;
- determination of performance standards;
- determination of fidelity requirements and evaluation of the sufficiency of fidelity present in the training device;
- determination of the sufficiency of instrumentation and procedures for collecting, reducing, analyzing, and displaying data;
- determination of the feasibility of using unique instructional features of the training device (e.g., ability to reinitialize itself after examinee makes a mistake or the ability to automate test scenarios and performance measurement) for testing; and
- determination of the acceptability of test program characteristics such as selection and training of examiners, frequency and scheduling (relative to other unit activities) of tests, and maintenance and logistics programs.

Each of these issues (plus many more) will have to be considered. For each issue, relevant decision criteria will have to be identified. Criteria can include subjective estimates of:

- dollar costs;
- non-dollar costs (e.g., safety);
- indirect costs associated with the diversion of limited unit resources (e.g., time, personnel, operational equipment, training devices);
- psychometric criteria such as reliability and validity;
- potential utility to users of test information;
- acceptability to examiners, examinees, and users of test information;

- practicality; and

- many other possible criteria that appear in the performance measurement literature such as non-reactivity, generalizability, and precision (Reference 44).

Criteria can be ranked in order of importance (overall or for each issue) and then each alternative measurement system (e.g., device versus operational system; or different device designs which may vary in fidelity and performance measurement capabilities) can be rated (formally or informally) on each criteria. Aggregation of ratings across criteria for each alternative can be accomplished using group judgment analysis techniques (e.g., Delphi, Nominal Group Techniques, multi-attribute utility analysis).

When there is disagreement between group members or when confidence in their judgment of a specific point is low, then plans can be made to address these points through empirical investigation - pilot testing and system effectiveness testing.

PILOT TESTING

As seen in the literature review, past empirical research on the use of training devices in proficiency testing has been mainly concerned with the selection of measure sets through assessment of their relative validities. Validity was inferred using several methods:

- simple correlation of performance in the training device with performance on the operational system;

- analysis (using inferential statistics or subjective analyses) of the consistency of individual differences across performances in both types of equipment;

- analysis (using inferential statistics or subjective analyses) of consistency of performance across levels of independent variables (e.g, as task difficulty is increased, do performance levels decrease a similar amount in both the device and the operational system);

- analysis of the relationship between automated measures of performance on the device and subjective analyses of performance on the device by instructors; and

- analysis of the ability of measures on the device to distinguish between masters and non-masters.

Although validity is important, pilot testing can also be used to gather other types of information. Ratings of potential test utility (e.g., accuracy, timeliness of feedback) can also be obtained from personnel who construct, administer, and use test information. Acceptance ratings can also be obtained from examiners, examinees, and users of test information. These ratings can provide insight into potential problems of implementing alternative test programs.

Experiments can also be set up to assess specific issues. For example, it may be better to deliver training devices to operational units with measurement "packages" (automated instrumentation and programmed procedures for proficiency testing). An experiment could be conducted comparing use of the training device for proficiency testing with a package versus use of the device without a measurement package (e.g., units would have to construct, administer and interpret tests using their own resources).

SYSTEM COST EFFECTIVENESS TESTING

Measurement system cost effectiveness testing is probably too expensive and time-consuming to be used to select alternative measurement systems. It is, however, needed as a check on the empirical utility of the selected measurement system.

It is recommended that longitudinal research programs be set up which measure changes in unit programs as a result of improved measurement techniques. Surveys on the use of training devices in training programs in operational units have revealed large discrepancies between the potential use of these devices (as intended by device and training program designers) and the actual use of these devices.

The actual patterns of implementation of devices in test programs in operational units will determine the actual utility of the devices for performance testing. If the devices are not used as they were designed to be used (due to lack of command emphasis, faulty maintenance, lack of acceptance, lack of training of examiners to use performance measurement capabilities, etc.), then test information will not be useful regardless of its original psychometric properties. Implementation must be assessed to insure useful performance measurement research for the Army.

SUMMARY

Substituting training devices for operational systems to assess operational readiness presents a complex problem. A systematic methodology that can serve as a model for all operational systems may solve the problem. To epitomize the methodology, however, a training device of an existing operational system has to be selected. This device should have training effectiveness analysis and cost effectiveness data readily available in order to identify and document:

- Existing or proposed measurement capabilities,
- Existing or proposed evaluation utilities, and
- Existing or proposed users.

The selection permits:

- the identification of both the users and uses within the U. S. Army evaluation programs (e.g., the EPMS at present, as well as, an aggregate of individual and collective performance data for unit readiness indices in future programs), and
- the determination of what to measure and how to design test programs to facilitate operational readiness assessment

through empirical investigations, viz., (a) pilot testing to determine whether and how the device will solve the Army's assessment needs and, if possible, (b) system cost effectiveness testing for longitudinal investigations of changes in measurement techniques for assessing operational effectiveness.

REFERENCES

1. Alluisi, E. A. Methodology in the use of synthetic tasks to assess complex performance. Human Factors, 1967, 9 (4), 375-384.

2. Alluisi, E. Performance measurement technology: Issues and answers. In Proceedings of the Symposium on Productivity Enhancement: Personnel Assessment in Navy Systems. San Diego, California: Naval Personnel Research and Development Center, October 1977.

3. American Airlines, Inc. Optimized flight crew training: a step toward safer operations. Fort Worth, Texas: American Airlines, Flight Training Academy, April 1969.

4. Angell, D., Sherer, J. W. & Berliner, D. C. Study of training performance evaluation techniques (NTDC 1449-1). Orlando, Florida: Naval Training Device Center, October 1964. (AD 609 605).

5. Bercos, J., Chiorini, J., Eakins, R., Lokie, A. & Stevens, W. Reserve component unit evaluation analysis: Volume I (Report prepared for the Department of the Army, Contract DAAG 39-75-C-0135). Falls Church, Virginia: Litton Mellonics Systems Development Division, October 1976.

6. Bergman, B. A. & Siegel, A. I. Training evaluation and student achievement measurement: A review of the literature (AFHRL-TR-72-3). Lowry Air Force Base, California: Air Force Human Resources Laboratory, January 1972. (AD 747 040)

7. Besnard, G.& Briggs, L. Measuring job proficiency by means of a performance test. In E.Fleishman (Ed.), Studies in personnel and industrial psychology. Homewood, Illinois: The Dorsey Press, 1967.

8. Blaiwes, A. S., Puig, J. A. & Regan, J. J. Transfer of training and the measurement of training effectiveness. Human Factors, 1973, 15 (6), 523-533.

9. Blanchard, R. E. Human performance and personnel resource data store design guidelines. Human Factors, 1975, 17, 25-34.

10. Braby, R., Henry, J. M., Parrish, W. F. & Swope, W. M. A technique for choosing cost-effective instructional delivery systems (TAEG Report No. 16). Orlando, Florida: Training Analysis and Evaluation Group, April 1975.

11. Caro, P. Aircraft simulators and pilot training. Human Factors, 1973, 15 (6), 502-509.

12. Caro, P. Some factors influencing Air Force simulator training effectiveness (HumRRO-TR-77-2). Alexandria, Virginia: Human Resources Research Organization, March 1977.

13. Charles, J. P. The simulator instructor - a readiness problem. Ninth NTEC/Industry Conference Proceedings (NAVTRAEQUIPCEN IH-276). Orlando, Florida: Naval Training Equipment Center, 1976, 211-215.

14. Chiles, W. D. Methodology in the assessment of complex performance: discussion and conclusions. Human Factors, 1967, 9 (4), 385-392.

15. Chiles, W. D. & West, G. Multiple task performance as a predictor of the potential of air traffic controller trainees: A follow-up study (No. 74-10). Washington, D. C.: FAA Office of Aviation Medicine, 1974.

16. Christensen, J. M. & Mills, R. G. What does the operator do in complex systems. Human Factors, 1967, 9 (4), 329-340.

17. Copperman, N. & Dorian, P. A. Using CAI to measure team readiness. Ninth NTEC/Industry Conference (NAVTRAEQUIPCEN IH-276). Orlando, Florida: Naval Training Equipment Center, 1976, 187-195.

18. Crawford, A. & Brock, J. Using simulators for performance measurement. In Proceedings of the Symposium on Productivity Enhancement: Personnel Assessment in Navy Systems. San Diego, California: Naval Personnel Research and Development Center, October 1977.

19. Crawford, M. P. Simulation in training and education (Professional Paper 40-67). Alexandria, Virginia: Human Resources Research Office, September 1967.

20. Cream, B. W., Eggemeier, F. T. & Klein, G. A. Behavioral data in the design of aircrew training devices. In Proceedings of the 19th Annual Meeting of the Human Factors Society. Dallas, Texas: Human Factors Society, 1975, 260-265.

21. Cronbach, L., & Gleser, G. Psychological tests and personnel decisions. Urbana, Illinois: University of Illinois Press, 1965.

22. Department of the Army. Field manual 100-5: Operations. Washington, D. C.: Author, July 1976.

23. Department of the Army. Pamphlet No. 350- (draft): SQT - guide for leaders. Washington, D. C.: Author, April 1977.

24. Department of the Army. Army regulation 220-1: Field organizations unit readiness reporting. Washington, D. C.: Author, March 1975.

25. Edwards, W., Guttentag, M., & Snapper, K. A decision-theoretic approach to evaluation research. In E. Struening and M. Guttentag (Eds.), Handbook of evaluation research (Vol. 1.) Beverly Hills: SAGE Publications, 1975.

26. Finley, D. L.,Rheinlander, T. W., Thompson, E. A. & Sullivan, D. J. Training effectiveness evaluation of Naval training devices Part I: A study of the effectiveness of a carrier air traffic control center training device (NAVTRAEQUIPCEN 70-C-0258-1). Orlando, Florida: Naval Training Equipment Center, August 1972.

27. Fitzpatrick, R. & Morrison, E. J. Performance and product evaluation. In R. L. Thorndike (Ed.), Educational measurements (2nd ed). Washington, D. C.: American Council on Education, 1971.

28. Gagne, R. M. Simulators. In R. Glaser (Ed.), Training research and education. Pittsburgh: University of Pittsburgh Press, 1962.

29. Glaser, R. & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.) Psychological principles in system development. New York: Holt, Rinehart and Winston, 1962.

30. Grodsky, M. A. The use of a full scale mission simulation for the assessment of complex operator performance. Human Factors, 1967, 9 (4), 341-348.

31. Hansen, D. N., Harris, D. A. & Ross, S. Flexilevel adaptive testing paradigm: validation in technical training (AFHRL-R-77-35(I)). Lowry AFB, Colorado: Technical Training Division, Air Force Human Resources Laboratory, July 1977.

32. Havens, C. B. Future academic training: A conservative projection of state of the art. Pensacola, Florida: Chief of Naval Air Training Command, 1975.

33. Hayes, J. F. & Wallis, M. R. ARTEP validation report. Alexandria, Virginia: U. S. Army Research Institute for the Behavioral and Social Sciences, in press.

34. Heymont, I. What is the army getting for its training dollar? Army, 1977, 27, (6), 34-38.

35. Hopkins, C. O. How much should you pay for that box? In Proceedings of the 19th Annual Meeting of the Human Factors Society. Dallas, Texas: Human Factors Society, 1975, i-vi.

36. Jeantheau, G. A. & Andersen, E. G. Training system use and effectiveness evaluation (NTDC 1743-1). Orlando, Florida: Naval Training Device Center, July 1966. (AD 640 423)

37. Jolly, O. & Caro, P. A determination of selected costs of flight and synthetic flight training. Fort Rucker, Alabama: Human Resources Research Organization, April, 1970.

38. Jones, R. A. The F-111D simulator can reduce cost and improve aircrew evaluation (Air War College Research Report No. 5959). Maxwell Air Force Base, Alabama: Air War College, Air University, United States Air Force, April, 1976. (AD 010 452)

39. Katz, M. Planning for performance measurement R&D: U. S. Army. In Proceedings of the Symposium on Productivity Enhancement: Personnel Assessment in Navy Systems. San Diego, California: Naval Personnel Research and Development Center, October 1977.

40. Knoop, P. A. Advanced instructional provision and automated performance measurement. Human Factors, 1973, 15 (6), 583-597.

41. Knoop, P. A. & Welde, W. L. Automated performance assessment in the T-37: a feasibility study (AFHRL TR 72-6). Wright-Patterson AFB, Ohio: Air Force Human Resources Laboratory, April 1973.

42. Koonce, J. M. Effects of ground based aircraft simulator motion conditions upon prediction of pilot proficiency (Technical Report ARL-74-5/AFOSR-74-3). Savo, Illinois: University of Illinois, Institute of Aviation, Aviation Research Laboratory, April 1974.

43. Mackie, R. R. Toward a criterion of training device acceptance. In Proceedings of the 19th Annual Meeting of the Human Factors Society. Dallas, Texas: Human Factors Society, 1975, 37-41.

44. Muckler, F. Selecting performance measures: "Objective" versus "subjective" measurement. Proceedings of the Symposium on Productivity Enhancement: Personnel Assessment in Navy Systems. San Diego, California: Naval Personnel Research and Development Center, October 1977.

45. Obermayer, R. W. & Muckler, F. A. Performance measurement in flight simulation studies (NSA CR-82). Washington, D. C.: National Aeronautics and Space Administration, July 1964.

46. Obermayer, R. W. & Vreuls, D. Combat-ready crew performance measurement system: phase IIIA crew performance measurement (AFHRL-TR-74-108 (IV)). Brooks Air Force Base, Texas: Headquarters Air Force Human Resources Laboratory (AFSC). December 1974. (AD B005 520L)

47. Obermayer, R. W. , Vreuls, D., Muckler, F., Conway, E. J. & Fitzgerald, J. Combat-ready crew performance measurement system study (Report prepared for the U. S. Air Force, Contract F41609-71-C-0008). Williams Air Force Base, Arizona: Flying Training Division, Air Force Human Resources Laboratory, May 1972.

48, Ontiveros, R. J. Effectiveness of a pilot ground trainer as a part-task instrument flight rules flight-checking device: Stage I (Report No. FAA-RD-75-72). Washington, D. C.: Federal Aviation Administration, Systems Research and Development Service, September 1975. (AD A015 722)

49. Ontiveros, R. J. Effectiveness of a pilot ground trainer as a part-task instrument flight rules flight-checking device: Stage II (Report No. FAA-RD-76-72). Atlantic City, New Jersey: Federal Aviation Administration, National Aviation Facilities Experimental Center, June 1976.

50. Osborn, W. C. Developing performance tests for training evaluation (HumRRO-PP-3-73). Alexandria, Virginia: Human Resources Research Organization, February 1973.

51. Povenmire, H. K. & Ballantine, K. M. Automated scoring of instrument flight checks. Ninth NTEC/Industry Conference Proceedings (NAVTRAEQUIPCEN IH-276). Orlando, Florida: Naval Training Equipment Center, 1976, 211-215.

52. Prophet, W. W. & Caro, P. W. Simulation and aircrew training and performance (HumRRO-PP-4-74). Alexandria, Virginia: Human Resources Research Organization, April 1974. (AD 780 688)

53. Prophet, W. W., Caro, P. W. & Hall, E. Some current issues in the design of flight training devices (HumRRO-PP-5-72). Alexandria, Virginia: Human Resources Research Organization, March 1972.

54. Rosen, M. H. & Behringer, R. D. Final report: M16AI rifle marksmanship training development. Springfield, Virginia: Washington Scientific Support Office, Mellonics Systems Development Division, September 1977.

55. Semple, C. Training effectiveness evaluation: Device 1D23, communications and navigation trainer (NAVTRAEQUIPCEN-72-C-0209-2). Orlando, Florida: Naval Training Equipment Center, November 1973.

56. Schum, D. A. Behavioral decision theory and man-machine systems. In K. B. De Greene (Ed.): Systems psychology. New York, New York: McGraw-Hill, 1970.

57. Shipley, B. D., Hagin, W. V. & Gerlach, V. S. Simplifying the measurement of complex skills in a training simulator. Ninth NTEC/Industry Conference Proceedings (NAVTRAEQUIPCEN IH-276). Orlando, Florida: Naval Training Equipment Center, 1976, 259-263.

58. Smode, A. F. Training device design: human factors requirements in the technical approach (NAVTRAEQUIPCEN 71-C-0013-1). Orlando, Florida: Naval Training Equipment Center, August 1972. (AD 754 802

59. Smode, A. F., Gruber, A. & Ely, J. H. Human factors technology in the design of simulators for operator training (NAVTRADEVCEN 1103-1). Orlando, Florida: Naval Training Device Center, December 1963. (AD 432 028)

60. Smode, A. F. & Hall, E. R. Translating information requirements into training device fidelity requirements. In Proceedings of the 19th Annual Meeting of the Human Factors Society. Dallas, Texas: Human Factors Society, 1975, 33-36.

61. SofTech, Inc. Task 3 report: Draft for ARPA and COTR review - The army training and evaluation system. Waltham, Massachusetts: Author, March 1977.

62. Stewart, S. R., Christie, C. I. & Jacobs, T. O. Performance correlates of the Dragon training equipment and the Dragon weapon system (report number NAVTRAEQUIPCEN N61339-74-C-0056-1). Alexandria, Virginia: Human Resources Research Organization, March 1974.

63. Sweezv, R., Chitwood, T., Easley, D. & Waite, B. Implications for Dragon gunnery training development. Washington, D. C: Litton Mellonics, Washington Scientific Support Office, September 1977.

64. Sweezy, R. & Pearlstein, R. Developing criterion-referenced tests. Reston, Virginia: Applied Science Associates, Inc., September 1974.

65. Sugarman, R. C., Johnson, S. L. & Hinton, W. M. SAT revisited - a critical post-examination of the systems approach to training. In Proceedings of the 19th Annual Meeting of the Human Factors Society. Dallas, Texas: Human Factors Society, 1975, 271-273.

66. Trans World Airlines. Flight simulator evaluation. Kansas City, Missouri: Trans World Airlines, Flight Operations Training Department, June 1969.

67. U. S. Army Training and Doctrine Command. Analyzing training effectiveness (TRADOC PAM 71-8). Fort Monroe, Virginia: Author, December 1975.

68. Valverde, H. H. A review of flight simulation transfer of training studies. Human Factors, 1973, 15 (6), 510-523.

69. Vineberg, R. & Taylor, E. Performance test development for skill qualifications testing: a manual (draft). Arlington, Virginia: U. S. Army Research Institute for the Behavioral and Social Sciences, August 1975.

70. Vreuls, D. & Woodridge, L. Aircrew performance measurement. In Proceedings of the Symposium on Productivity Enhancement: Personnel Assessment in Navy Systems. San Diego, California: Naval Personnel Research and Development Center, October 1977.

71. Vreuls, D., Woodridge, A. L., Obermayer, R. W., Johnson, R. M., Norman, D. A. & Goldstein, I. Development and evaluation of trainee performance measures in an automated instrument flight maneuvers trainee (NAVTRAEQUIPCEN 74-C-0063-1). Orlando, Florida: Naval Training Equipment Center, October 1975. (AD A024 517)

72. Waag, W. L., Eddowes, E. E., Fuller, J. H. & Fuller, R. R. Advanced simulation in undergraduate pilot training (ASUPT) automated objective performance measurement system. Catalog of Selected Documents in Psychology, 1975 (Fal), 5, 358.

73. Waag, W. L., Eddowes, E. E., Fuller, J. H. & Fuller, R. R. ASUPT automated objective performance measurement system (AFHRL-TR-75-3). Williams Air Force Base, Arizona: Air Force Human Resources Laboratory, March 1975. (AD 014 799)

74. Weitzman, D. O., Fineberg, M., Ozkaptan, H. & Compton, G. L. Evaluation of the synthetic flight training system (device 2 B24) for maintaining IFR proficiency among experienced pilots. Ninth NTEC/Industry Conference Proceedings (NAVTRAEQUIPCEN IH-276). Orlando, Florida: Naval Training Equipment Center, 1976, 63-68

75. Williges, B. H., Roscoe, S. N. & Williges, R. C. Synthetic flight training revisited. Human Factors, 1973, 15 (6), 543-560.

76. Williges, R. Automation of performance measurement. Proceedings of the Symposium on Productivity Enhancement: Personnel Assessment in Navy Systems. San Diego, California: Naval Personnel Research and Development Center, October 1977.

77. Xyzyx Information Corporation. Evaluation of the CATCC team trainer as a performance qualification instrument (NADC Contract No. N62269-73-C-0109). Washington, D. C.: Naval Air Systems Command, January 1973.

APPENDIX A

PERSONNEL CONTACTED DURING STUDY

Capt. G. McCulloch
Manager, Flight Training Division
Stapleton Training Center
Denver, Colorado

Capt. M. Cavenaugh
Director, Flight Standards and
Procedures
Stapleton Training Center
Denver, Colorado

Mr. G. Cohen
Flight Standards Branch
Federal Aviation Administration
Washington, D.C.

Dr. E. Eddows
Flying Training Division
U. S. Air Force Human Resources
Laboratory
Williams AFB, Arizona

Col. D. Berjon
Military Airlift Command
Offutt AFB, Nebraska

LCDR. J. Ashburn
Naval Air Systems Command
Washington, D.C.

Mr. J. Puig
Naval Training Equipment Center
Orlando, Florida

Mr. R. Browning
Training Analysis and Evaluation
Group
U. S. Navy
Orlando, Florida

Maj. Monday
Individual Training Evaluation
Directorate
Test Development and Research
Division
U. S. Army Training Support Center
Fort Eustis, Virginia

CPT Barlow
SM/SQT Branch
Design Division
Directorate of Training Development
U. S. Army Infantry School
Fort Benning, Georgia

Dr. P. Caro
Seville Research Corporation
400 Plaza Building
Pensacola, Florida

APPENDIX B

CANDIDATE COST ELEMENTS

Acquisition Cost Elements

Government Procurement Costs (*e.g.*, requirements, specification, negotiation, contract monitoring costs)
Basic Device Hardware
Separately Identifiable Features (*e.g.*, motion, visual system, instructor station)
Computer Complex and Peripherals
Test Equipment
Computer Software
Training Package Software
Maintenance and Test Equipment
Aircraft Data
Aircraft Parts for Simulation
Manufacturer Supplied Maintenance Training
R&D
T&E
Facilities Construction
Packaging and Shipping
Installation (material and labor)
Acceptance Testing
Manufacturer Field Representatives (technical support, installation)
Military Personnel (management, acceptance team)
Initial Spares
Standard Documentation
Aircraft Cost (total cost to purchase)

Operating and Support Cost Elements

Base Level-Simulation

- Buildings and Facilities
  - Building depreciation
  - Security
  - Utilities/energy
  - Janitorial service
  - Maintenance
  - Office equipment
- Simulator Depreciation
- Personnel
  - Training management
  - Management support
  - Instructors
  - Simulator Operators
  - Students
  - Maintenance personnel (government, contractor)
  - Contractor support

Simulator Modification Costs (amortized) (materials and labor, government and contractor)

Maintenance Materials and Parts (replenishment - expendable and repairable)

Base Level - Training Aircraft

Buildings and Facilities
- Building depreciation (hangers)
- Utilities/energy
- Maintenance

Aircraft Costs
- Depreciation
- Materials, parts & lubricants (replenishment)
- Fuel
- Modification costs (amortized)

Personnel
- Training and base management
- Instructors
- Students
- Maintainers
- Support - (security, facilities maintenance, outside contractors, etc.)

Depot Level (Logistic Centers)

Buildings and Facilities

Maintenance Equipment Depreciation

Procurement Costs, Materials & Parts

Maintenance Personnel

Management Personnel

Base Operating Support Costs

Distribution/Shipping Cost

Other Transportation Costs

Personnel

Personnel Replacement Costs
- Recruiting
- Training Costs
  - Central Instructor School
  - Instructor under training, unit level
  - Maintainer - basic, specialty
  - Simulator operator
- PCS relocation

Personnel Support Costs
- Medical
- Administrative
- Material costs